

**Foal Immunodeficiency Syndrome:
Identification of the causal mutation**

**Thesis submitted in accordance with the
requirements of the University of Liverpool for
the degree of Doctor in Philosophy**

by

Laura Yana Fox-Clipsham, BSc

November 2011

Contents

	Page
Contents	3
Abstract	9
Acknowledgments	10
<u>Chapter 1: General Introduction</u>	
The Fell and Dales Pony	11
Fell Pony Syndrome (Foal Immunodeficiency Syndrome)	12
Clinical Presentation of Foal Immunodeficiency Syndrome	13
Clinical pathology of Foal Immunodeficiency Syndrome	17
Epidemiology of Foal Immunodeficiency Syndrome	20
Genealogy of Foal Immunodeficiency Syndrome	20
Comparable disease in the horse	23
Hereditary disorders of the horse which cause anaemia	23
Immunodeficiencies in the horse with a known genetic aetiology	23
Human genetic disorders which cause immunodeficiency and anaemia	28
Summary of Foal Immunodeficiency Syndrome	30
Investigational summary	32
<u>Chapter 2: Mapping an Associated Disease Locus Using Microsatellite Markers</u>	
Summary	33
2.1 Introduction	33
Disease mapping with microsatellite markers	33
Basic principles of genotyping microsatellites	34
Mapping traits using linkage mapping	35

Homozygosity mapping	37
Objectives and aims of the FIS whole-genome microsatellite mapping approach	39
2.2 Materials and Methods	40
Animals and Samples	40
DNA extraction from tissue and blood samples	41
Calculation of power to detect significant linkage	41
Microsatellite genome scan	41
Statistical analysis of the genotyping data	43
2.3 Results	44
Two-point linkage analysis results	47
Results of the homozygosity mapping	47
Multi-point linkage analysis	49
2.4 Discussion	52
<u>Chapter 3: Confirming the associated chromosome using SNP GWAS</u>	
Summary	55
3.1 Introduction	55
Recent advances in the equine genome	55
Association studies for detecting disease loci	57
Analysis of genetic association data – considering the study design	58
The basic principles of performing GWAS using the Illumina Beadchip	61
Objectives of the FIS disease mapping association study	63
3.2. Materials and Methods	64
Animals and Samples	64
DNA extraction from tissue and blood samples	65
Single Nucleotide Polymorphism Genotyping using the Beadchip	67

3.3. Results	71
Additional sample collection for this investigation	71
SNP genotyping analysis	76
Assessing population stratification	78
Genome-wide association mapping	80
Linkage disequilibrium of the FIS-associated region and haplotype based association test	85
3.4. Discussion	89
<u>Chapter 4: Fine mapping the critical region and interrogation of candidate genes</u>	
Summary	92
4.1. Introduction	92
Fine-mapping studies to narrow the critical chromosomal region	92
Selection of genetic markers for fine-mapping critical intervals in the horse	93
Selection and interrogation of candidate genes	94
Aims and Objectives of narrowing the critical chromosomal region and interrogation of potential candidate genes	96
4.2 Materials and methods	97
Animals and Samples	97
Fine-structure mapping	98
Examination of candidate genes in an attempt to identify the causal mutation	101
4.3 Results	104
Fine-mapping the homozygous critical region	104
Interrogation of the critical region – searching for candidate genes	107
4.4 Discussion	111

Chapter 5: Re-sequencing the critical interval to identify the causal variant

Summary	113
5.1. Introduction	113
Recent advances in genetics: The impact of next-generation sequencing	113
Sequencing using the Roche FLX Titanium series: Pyrosequencing technology	116
Selective re-sequencing: Capturing the target region	117
Analysis of next-generation sequencing data	118
Objectives of re-sequencing the FIS critical interval to identify the causal mutation	122
5.2. Materials and Methods	123
Animals and Samples	123
DNA extraction from tissue and blood samples	123
Designing the sequence capture array	124
Capturing the target sequence and sequencing the libraries	124
Analysis of the sequencing data	127
5.3. Results	133
Sequence capture, sequencing and mapping the reads	133
Refining the homozygous critical haplotype	134
Investigating large scale rearrangements	140
Identification and interrogation of sequence variants in the critical region	142
5.4. Discussion	145

Chapter 6: Population studies: Estimating the prevalence of FIS-carriers

Summary	150
6.1 Introduction	150
A population screen of the Fell and Dales breeds to estimate the prevalence of FIS-carriers	150
Population screen to assess the spread of FIS into other breeds	151

Pedigree analysis of Foal Immunodeficiency Syndrome affected foal	152
Aims and objectives of the FIS population studies	153
6.2 Materials and Methods	154
Population screen	154
Sample processing	156
Pedigree analysis of FIS-affected foals	159
6.3 Results	160
Population screen	160
Pedigree analysis of FIS-affected foals	161
6.4 Discussion	164
<u>Chapter 7: A pilot study to investigate transcriptional changes in FIS-affected foals</u>	
Summary	169
7.1 Introduction	169
Investigating global transcriptional changes	169
FIS and anticipated transcriptional changes	171
Why perform a pilot study?	172
The experimental design: A pilot study to investigate global transcriptional changes in FIS-affected foals	173
7.2 Materials and Methods	176
Animals and sample collection	176
Total RNA extraction from bone-marrow samples	177
RNA-Sequencing using the Illumina Genome Analyser II System	178
Analysis of RNA-Seq data: Detecting differentially expressed genes	180
Analysis of molecular interactions	183
7.3 Results	185
Animals and Samples: Selection of samples for RNA-Seq	185

Mapping the reads	186
Identifying differentially expressed genes	187
Pathway analysis using KEGG Mapper	199
Pathway analysis using Ingenuity	202
7.4 Discussion	209
<u>Chapter 8: General discussion and Future studies</u>	
Mapping the associated loci using genetic markers	216
Next-generation sequencing – A revolutionary tool for mutation mining	217
The sodium/glucose co-transporter family	218
Evaluation of the sodium/myo-inositol co-transporter as the causal mutation of FIS: Structural organisation and functional analysis	220
Breeding management to reduce carrier prevalence	227
Pathological and biological effects of FIS	229
Conclusion and final remarks	230
Appendices	231
Abbreviations, Buffers and Reagents	261
Bibliography	262
Supporting Articles	279
1. Immunodeficiency/anaemia syndrome in a Dales pony	
2. Identification of a Mutation in SLC5A3 Related to Fatal Foal Immunodeficiency Syndrome in the Fell and Dales Pony	



Abstract

Foal Immunodeficiency Syndrome (FIS), is a disease that affects both Fell and Dales Ponies. FIS results in a profound anaemia and a severe deficiency in the number of circulating B-lymphocytes. Consequently, FIS-affected foals begin to lose condition and suffer from multiple opportunistic infections; FIS is eventually fatal. Pedigree analysis, incorporating the knowledge of FIS-affected individuals, suggested that FIS is a genetic disorder, with an autosomal recessive mode of inheritance. Further, analysis of the Fell and Dales Pony stud book revealed a common founder stallion, in the maternal and paternal lineage of all Fell and Dales FIS-affected individuals. Based on this, the primary aim of this investigation was to characterise the genetic lesion responsible for FIS, and subsequently develop a diagnostic test which could be used to identify asymptomatic carriers.

Two approaches were taken in this study to definitively map the FIS locus; a microsatellite whole-genome scan, and a genome-wide association study. Linkage analysis and homozygosity mapping of the microsatellite marker data revealed a single locus which showed significant linkage to FIS. This was then further supported with the genome-wide association study, using the EquineSNP50 Beadchip, which identified the same locus with a significant disease association. After additional fine-mapping of the associated region, four plausible candidate genes were identified and subsequently investigated, although none revealed the causative mutation. Therefore, the entire FIS critical interval was sequenced using next-generation re-sequencing. This led to the identification of a non-synonymous single nucleotide polymorphism, in the single exon of the sodium/myo-inositol cotransporter gene (*SLC5A3*), which is highly associated with the FIS phenotype. This gene plays a crucial role in the osmoregulation of tissues, a process which has been shown to be extremely important in the development of lymphoid tissues, lymphocytes, peripheral nerves and during early embryonic development. Further functional studies are now required to assess the functional consequences of this mutation.

The identification of this mutation has led to the development of a diagnostic test, which is not only used to identify asymptomatic carriers of FIS, but also to definitively diagnose foals which are suspected FIS-affected. Additionally, this test has been used to perform a population screen, to assess the prevalence of the FIS mutation and investigate possible transfer into other equine breeds. This revealed that carrier prevalence is approximately 40-50% in the Fell Pony and 10-20% in the Dales Pony. Further, this confirmed that the FIS mutation had transferred into the Coloured pony population.

Global transcriptional changes in FIS-affected foals were evaluated as part of a pilot study. This revealed significant disruption of multiple pathways, including those responsible for the haematological system and its development, tissue development and cell growth. Due to the severe clinical presentation of the FIS-affected foals used in this study, this data provides limited information on primary consequences of the causal variant. Rather it provides a snapshot of the transcriptional changes associated with the downstream effects of the FIS-causal variant and the multiple pathological effects associated with this.

Acknowledgments

Initially, I would like to thank my supervisors, Dr. June Swinburne and Prof. Stuart Carter, without whom this project would not have been possible. They have both been tremendously supportive, with special thanks to June for having the patience of a saint with me – putting up with me on a daily basis with my umpteen questions.

I would like to thank all the veterinarians and breeders who have supported this work by helping me to collect valuable samples – special thanks to Paul May and Thomas Capstick. Thanks also to all those at Liverpool University who have helped me with this project, especially Prof. Derek Knottenbelt for his help with collecting bone marrow samples, Fernando Malalana for his support with the first case of FIS in a Dales foal, and Di Isherwood for spending time rummaging through freezers with me.

Thanks to all those at the Animal Health Trust who have helped me along this journey, especially Louise Downs, Tony Blunden, Debs Flack and Netty Flindall. Further I would like to thank Prof. William Ollier for his input and inspiration, and everyone at the Centre for Genomic Research at the University of Liverpool, especially Ian Goodhead and Prof. Neil Hall.

Special thanks to The Horse Trust for their financial support, as without this the project would not have been possible. Thanks must also go to the Fell and Dales Pony Societies, who have always been very supportive of this work.

Finally, but most importantly, love to my husband, who has been behind me every step of the way.

Chapter 1

General Introduction

	Page
The Fell and Dales Pony	11
Fell Pony Syndrome (Foal Immunodeficiency Syndrome)	12
Clinical Presentation of Foal Immunodeficiency Syndrome	13
Clinical pathology of Foal Immunodeficiency Syndrome	17
Epidemiology of Foal Immunodeficiency Syndrome	20
Genealogy of Foal Immunodeficiency Syndrome	20
Comparable disease in the horse	23
Hereditary disorders of the horse which cause anaemia	23
Immunodeficiencies in the horse with a known genetic aetiology	23
Human genetic disorders which cause immunodeficiency and anaemia	28
Summary of Foal Immunodeficiency Syndrome	30
Investigational summary	32

The Fell and Dales Pony

Fell Ponies are an old breed of horse and are native to the North of England and most commonly found roaming freely in the region of the Cumbrian Fells. The first record of Fell Ponies was documented in the late 19th century, when pedigree records began to be kept, and then in 1922 the Fell Pony Society formed, to 'keep pure the old pony breed'. Traditionally the Fell pony has been used as a working pony and in the early 20th century, the mining community capitalised on this hardy and sure-footed pony breed. Miners used them underground where possible and above ground for transporting coal in the north-east of England and transporting copper, iron and lead ores from mines in the north-west of England to the smelting works. Although a large proportion of the breeding stock are still farmed in semi-feral herds on the Cumbrian Fell, the Fell Pony has become established as a popular recreational pony and can be found throughout the UK, across Europe and in regions of Canada and the United States of America.

The Dales Pony is native to the upper Dales and Eastern slopes of the Pennines, ranging from the High Peak in Derbyshire to the Cheviot Hills near the Scottish Border. Pedigree records for the Dales Pony are fairly good, with many of the Dales Ponies alive today tracing back to the foundation sire which was born in 1755. In 1916 the Dales Pony Improvement Society was formed and the Dales Pony stud book opened, keeping official records of all offspring. Today, the UK Dales Pony Society accepts registrations in the stud book provided the animal has three generations of recorded breeding on both sides. Additionally, Dales Ponies are categorised based on type, being registered as either section A, B or C. Traditionally used in the lead mining industry, the Dales Pony became renowned for great strength, endurance, and the ability to quickly cover rough terrain. Today Dales Ponies are popular riding and driving ponies, and make excellent family ponies due to their soft nature and versatility.

Fell Pony Syndrome

Over the past 15 years, the Fell Pony has suffered from a severe immunodeficiency and anaemia syndrome, formally known as Fell Pony Syndrome (FPS), which results in the loss of foals (Scholes et al., 1998). Personal communications with breeders and veterinarians in Cumbria has however revealed that anecdotally, the first loss of Fell Pony foals from FPS was probably in the late 1960s. Although difficult to estimate due to the natural husbandry of the Fell Pony, it was estimated that approximately 15-25% of the foals born annually were affected by this disease (Bell et al., 2001). Once identified, the disease was considered a major health and welfare issue for the Fell Pony breed, which led to research into the underlying pathogenesis and disease aetiology. It was through this research, primarily from analyses of the Fell pony stud books, that a genetic aetiology was identified as the probable cause. Due to the frequent out-breeding of the Fell Pony with other pony breeds, in particular the Dales Pony, it was considered highly likely that the genetic mutation responsible for this disease had transferred into other pony breeds, and would eventually give rise to affected foals in these other pony populations. This was confirmed when, in 2008, the first case of FPS was reported in a Dales Pony foal with three generations of Dales Pony pedigree on both the maternal and paternal side (Fox-Clipsham et al., 2009). At this time it was deemed appropriate to remove the breed specific attachment from the disease name and in November 2009 Fell Pony Syndrome was re-named Foal Immunodeficiency Syndrome (FIS).

Both Fell and Dales pony breeds are registered with the Rare Breeds Survival Trust due to their limited breeding stock. The threat of a disease which results in the loss of foals was devastating to these breeds, which were already deemed 'at risk'. Immediate action was therefore required to help preserve these breeds and prevent the untold suffering of affected foals. This led to the current study into the genetic basis of FIS.

Clinical Presentation of Foal Immunodeficiency Syndrome

Affected foals are apparently normal at birth but within the first few weeks of life, usually within four weeks, the foals begin to lose condition (Fig. 1.1) and suffer from multiple clinical signs which are associated with the syndrome (Scholes et al., 1998). Foals often present with a cough and chronic watery diarrhoea (Fig. 1.2C) which initially responds to supportive therapy and treatment but later becomes unresponsive and persistent. Pale mucous membranes (Fig. 1.2A), which are consistent with a progressive profound anaemia, are apparent in all affected foals. Other clinical signs include weight loss, nasal discharge (Fig. 1.2D), dry unkempt coats, dull demeanour, hypersalivation (Fig. 1.2B), frequent chewing movements and failure to suckle. Both the frequent chewing and hypersalivation are associated with ulceration and pseudomembranous coating of the tongue. Antibiotics and anthelmintic drugs have been proposed as beneficial in the treatment of syndrome foals (Richards et al., 2000), but despite a wide range of treatments and supportive therapies, foals die or are euthanized on the basis of lethargy, severe anaemia and persistent infections before 16 weeks of age. There have been no validated cases of foals surviving this disease.



Figure 1.1: *Clinical signs of Foal Immunodeficiency Syndrome. Affected foals lose condition and present with multiple clinical signs associated with immunodeficiency and anaemia. Here an FIS-affected foal presents with a poor demeanour, signs of dehydration and an unkempt coat. Photograph courtesy of Professor D. Knottenbelt.*

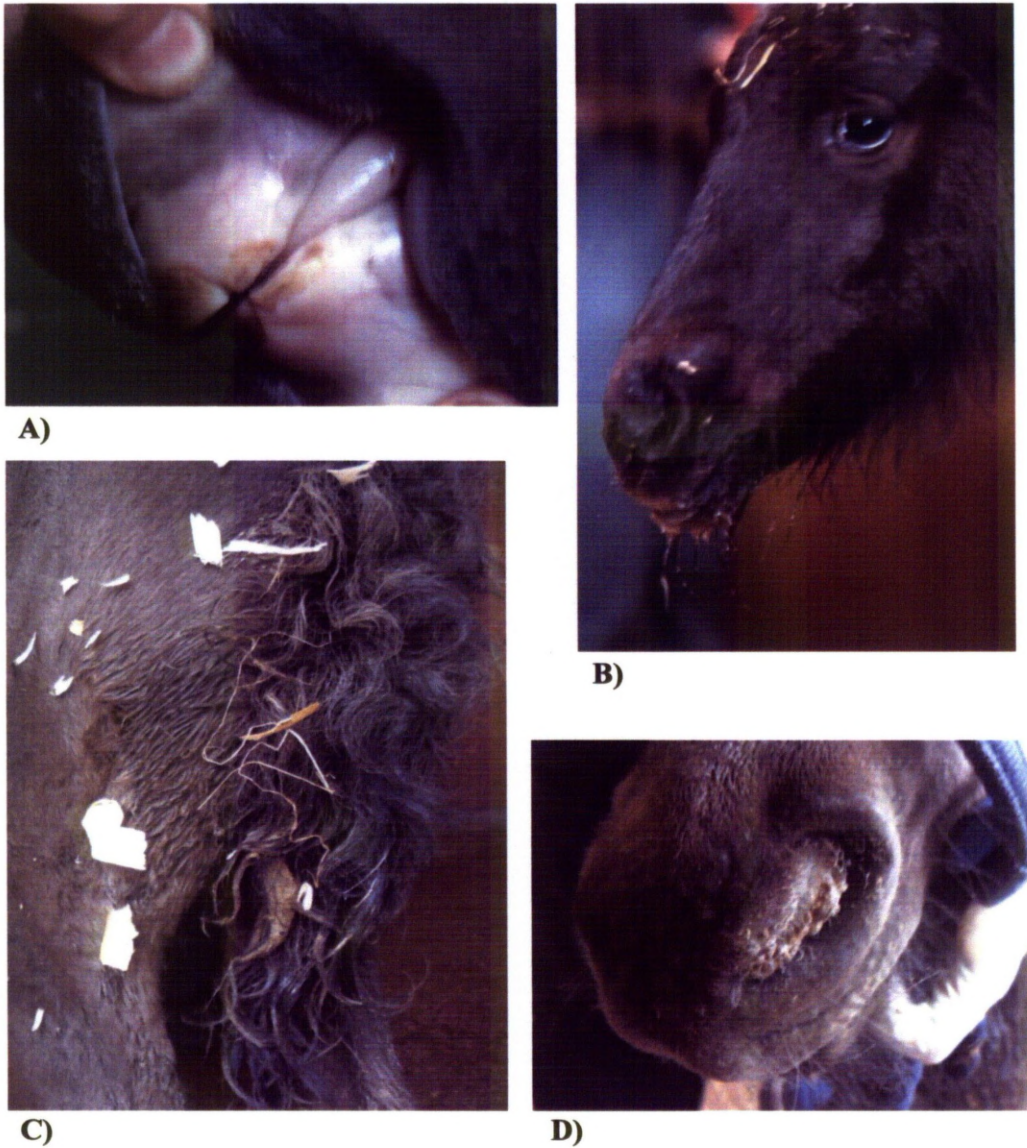


Figure 1.2: *Clinical features associated with Foal Immunodeficiency Syndrome. A) Pale mucous membranes consistent with severe anaemia. B) Foal exhibiting excessive salivation. C) Evidence of persistent watery, yellow coloured diarrhoea. D) Bilateral nasal discharge in an affected foal, which also had thoracic sounds.*

Haematological, immunoglobulin and peripheral B-lymphocyte analysis of FIS-affected foals

A severe progressive profound anaemia is found consistent with all FIS-affected foals ($P < 0.0001$) (Dixon et al., 2000) (Fig. 1.3), with Packed Cell Volumes (PCVs) as low as 4% recorded in live affected foals (P.May, unpublished observations). The anaemia is not associated with any blood loss or haemolysis and Coombs' test is negative showing that the anaemia is not related to an autoimmune reaction. The anaemia is usually normocytic and normochromic.

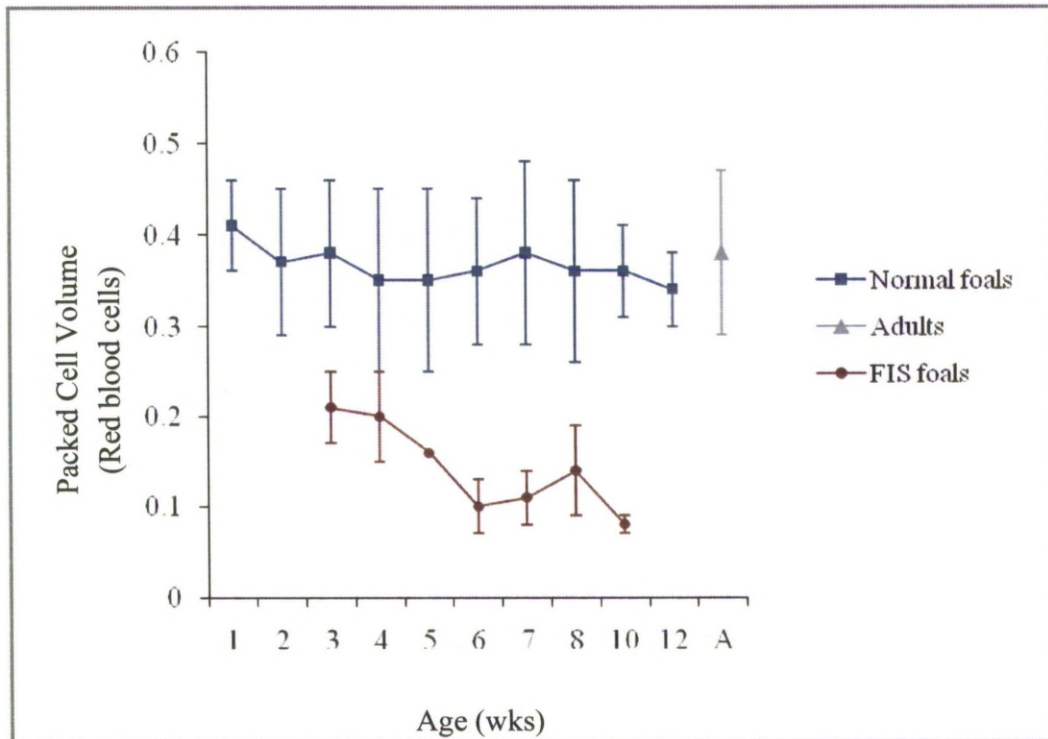


Figure 1.3: *Profound anaemia in Fell pony foals with FIS, which progresses with age. Data from normal foals and adults ponies is provided as a comparison. PCV Mean (\pm 1SD) PCV v age in weeks. Reproduced with permission from Thomas (2003).*

Flow cytometric analysis of peripheral blood samples revealed that FIS-affected foals have no significant alteration (compared to normal age-matched Fell pony foals) in circulating T-lymphocyte numbers and also display unaltered numbers of

CD4⁺ and CD8⁺ T-cells (Bell et al., 2001). Lymphopenia is a primary characteristic of FIS but Thomas (2003) showed that this is the result of a B-lymphocyte deficiency rather than a combined or T-lymphocyte deficiency. This characteristic forms a distinct difference between FIS and Severe Combined Immunodeficiency (SCID) which affects Arabian foals, as SCID foals have severely reduced numbers of both T and B-lymphocytes (McGuire and Poppie, 1973). The severe B-lymphocyte deficiency ($P < 0.001$) in FIS-affected foals was demonstrated by flow cytometric analysis of peripheral blood lymphocytes from diseased and normal ponies by Thomas (2003) (Fig. 1.4). This work also provided evidence that pre-FIS foals, that is, FIS foals sampled prior to showing clinical signs, also have significantly lower numbers of B-lymphocytes when compared to age matched controls (Fig. 1.4).

Consistent with this profound lack of B-lymphocytes, FIS-affected foals also have a significant reduction in immunoglobulin levels compared to age matched controls (Thomas et al., 2005). Reductions in all immunoglobulin classes are observed; IgM ($P < 0.001$), IgGa ($P < 0.001$), IgGb ($P < 0.005$) and IgG(T) ($P < 0.02$). The timing of these reductions coincides with the natural drop in maternally derived antibodies (derived from the mare's colostrum soon after birth). In normal foals this would be the point at which B-lymphocytes would start to secrete their own immunoglobulins, to take over from maternally derived antibodies. These data clearly indicate that affected foals are unable to produce their own immunoglobulins and so by the time maternally-derived immunity has diminished, affected foals have very few, if any, serum immunoglobulins and therefore succumb quickly to infections. In contrast to this study, Scholes *et al.* (1998) observed that the levels of IgM and IgG subisotypes were not significantly different between affected and unaffected foals. However, only two animals were used in this study, one of which was sampled at 24 days so maternal antibodies may have contributed to the absolute levels.

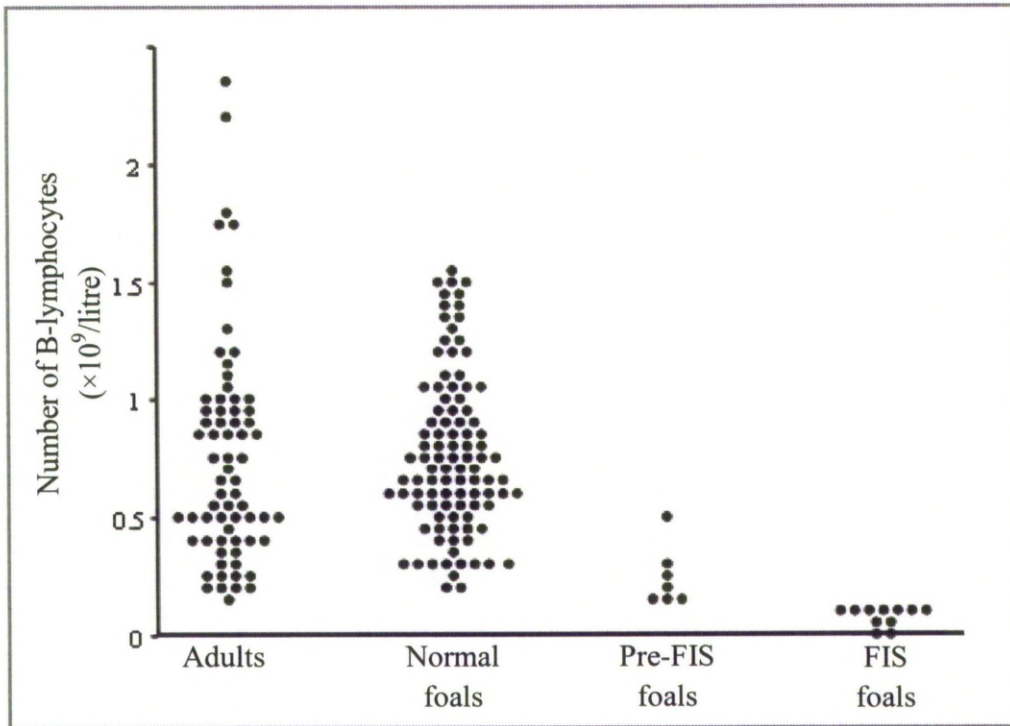


Figure 1.4: Numbers of B-lymphocytes in adult Fell Ponies, healthy Fell Pony foals, foals sampled at the onset of infection that were not clinically anaemic (pre-FIS) and foals displaying clinical signs of FIS. Reproduced with permission from Thomas (2003).

Clinical pathology of Foal Immunodeficiency Syndrome

Due to the lack of a definitive test for FIS-affected foals, those animals suspected as being affected are commonly euthanized to prevent suffering. Antemortem diagnosis of FIS-affected foals can be challenging due to the non-specific clinical signs. Usually diagnosis is made on clinical history, presentation of the foal and a severe progressive anaemia, which is a strong disease indicator. A further aid to antemortem diagnosis is flow cytometric examination of peripheral blood to determine the B-lymphocyte population. B-lymphocyte numbers in FIS-affected foals are significantly lower than in healthy age matched controls, so this test supports diagnosis (Thomas et al., 2005). Even though cytometric analysis acts as a good disease indicator, this test is rarely used in practice, because it is not a routine

laboratory test and it requires very specialised equipment and rapid testing of fresh blood samples. Although diagnosis of suspected FIS-affected foals cannot be confirmed definitively antemortem, it can be confirmed post-mortem based on histological examination, small lymphoid organs and a reduced number of erythrocyte precursors in the bone marrow (Thomas, 2003).

Necropsy findings in FIS-affected foals

Common findings at post-mortem examination included glossal hyperkeratosis (Fig. 1.5A), typhlocolitis, intestinal cryptosporidiosis (Fig. 1.5C), enteritis, pancreatitis and bronchopneumonia. Many of those infectious agents isolated from lesions, such as intestinal *Cryptosporidia sp.*, adenovirus and *Candida sp.* are opportunistic infections rarely seen in animals other than those with severely compromised immune systems.

Lymphoid organs also show evidence of immunocompromise. FIS-affected foals have a significantly reduced thymus compared to age matched controls, with small thymic lobules with no demarcation between the cortex and the medulla. Lymph nodes are also reduced in size, with drainage lymph nodes displaying evidence of inflammation. Histopathological examination of secondary lymphoid organs reveals a paucity of germinal centres (Fig. 1.5 A and B) and a marked reduction in the number of B-lymphocytes. Furthermore, the spleen is deficient in lymphoid cells, has sparse or absent germinal centres and no stromal support (Bell et al., 2001).

Bone marrow examination of affected foals is consistent with peripheral blood haematology, showing a significant reduction of erythroid precursors. Myeloid-erythroid ratios of 21-62:1 have been reported in FIS-affected foals (Jelinek et al., 2006, Bell et al., 2001), which is significantly elevated compared to normal myeloid-erythroid ratios of 0.5-1.5:1 (Jain, 1993). Femoral marrow is typically uniformly pink and is hypercellular.

Changes to the peripheral ganglia have also been noted in a number of FIS-affected foals (Scholes et al., 1998); reduced numbers of chromatolytic neurons and occasional axonal spheroids, with chromatolysis usually being central but rarely complete. However, these observations were not reported in a later study by Richards *et al.* (2000), who identified no lesions of the autonomic neurons.

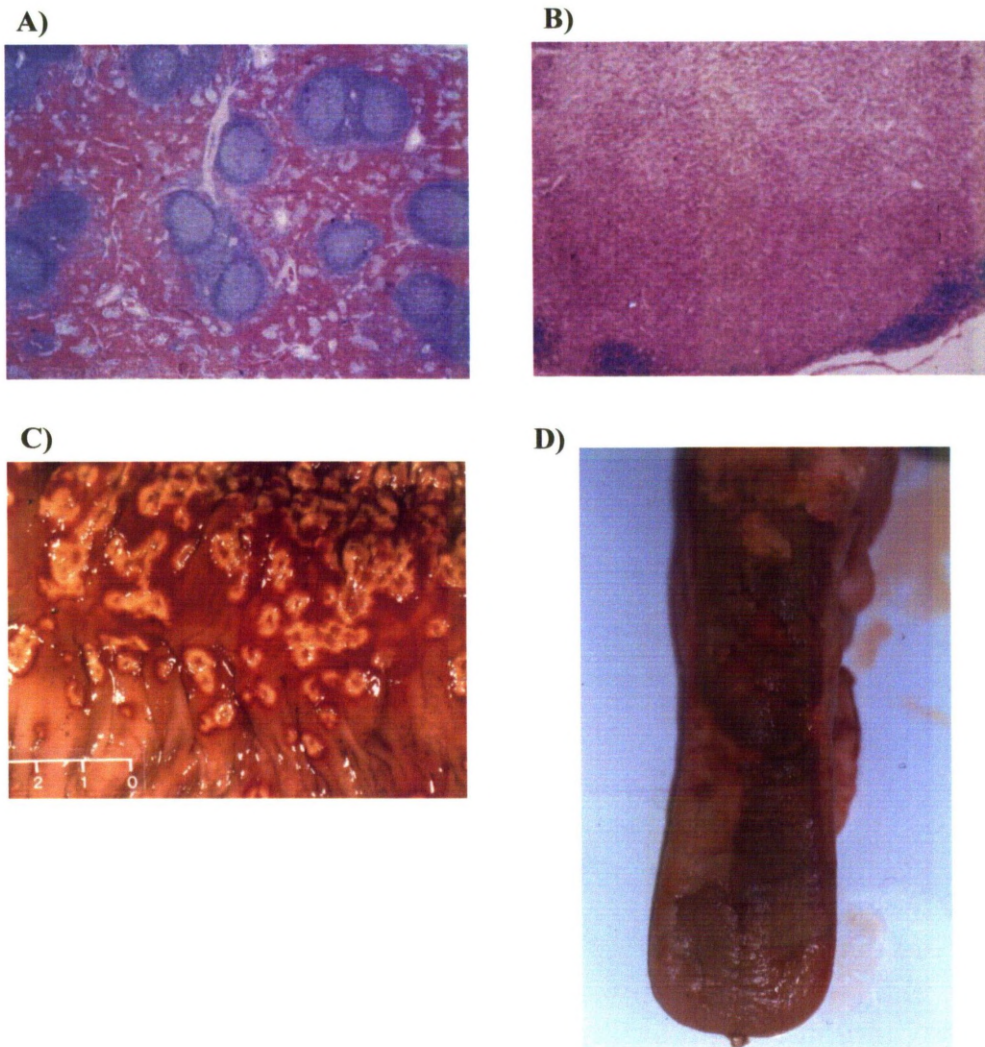


Figure 1.5: *Typical necropsy findings of FIS-affected foals. A) Normal spleen shows multiple basophilic germinal centres. B) Spleen from FIS-affected foal shows severe paucity of germinal centres and a loss of lymphoid tissue. C) Severe intestinal inflammation and ulceration. D) Ulceration of the tongue, with a pseudomembranous coating. Images reproduced from Thomas (2003).*

Epidemiology of Foal Immunodeficiency Syndrome

Foal Immunodeficiency Syndrome has been reported in Fell Ponies throughout the UK, across Europe and in the United States of America, i.e. in all Fell Pony populations. To date, this disease has only been confirmed in a single Dales Pony foal that was from Northern England. There is anecdotal, but unconfirmed reports of a similar disease occurring sporadically in the Coloured Pony, particularly in those herds that are known to have interbred with either the Fell or Dales.

Due to the natural husbandry of the Fell Pony breed, with many herds still roaming freely on the Cumbrian Fells, it is very difficult to provide an accurate estimate of disease prevalence. The only report of disease prevalence was in 2001, when it was estimated that the dramatic decline (15-25%) of Fell Pony foal registrations was primarily due to the loss of foals from FIS (Bell et al., 2001). The incidence of FIS within the large majority of breeding herds has been estimated by Cumbrian based veterinarians at <15%. However, due to the sporadic nature of this disease and given the fact that many Fell Ponies foal on the Cumbrian Fell with no human intervention, it is unlikely that veterinary involvement is sought in all cases.

Genealogy of Foal Immunodeficiency Syndrome

The effective population size of the Fell Pony declined dramatically around the time of the Second World War, when the introduction of tractors and railways made many ponies redundant, many of which were sent to slaughter. Around this time, the enclosure system was also introduced, in which mares would be taken to the stallion for covering, rather than the more traditional method where stallions moved from farm to farm over a breeding system. Enclosure enabled popular stallions to cover many more mares during any one season. Analysis of the Fell Pony stud book (Thomas, 2003) revealed that the majority of stallions which were used at stud in the 1970s, traced back to 6 stallions born in the 1930s and 1940s. Furthermore, 93% of the Fell Pony foal registrations made in 1998 trace back to three of these popular stallions. Both the dramatic reduction in numbers and the use of the enclosure

system would have resulted in a loss of genetic variation within the breed, forming a genetic bottleneck.

The Dales Pony, which is a close relative of the Fell Pony, was commonly crossed with the Fell Pony at various times during the 20th Century. The Dales was known for its endurance and sure footedness, and used by the army in the Second World War. Over 200 Dales Ponies were taken by the army for use in the war, very few of which returned. The Second World War nearly destroyed the breed and it was at this time that the 'inspection system' was introduced in an attempt to preserve the Dales. The inspection system introduced a grading system in which mares of suitable 'type' which were not registered as Dales Ponies, such as Fell Pony mares, could mate with a Dales stallion and the offspring registered as Dales. In addition, Fell stallions could also be registered in the Dales studbook. In 1971, the number of Dales Pony registrations had risen gradually and the inspection system was closed. Unlike the Fell Pony Society, the Dales Society attempted to limit the effects of the dramatic reduction in numbers by introducing new blood lines from the Fell Pony. However, those Fell Ponies used were selected based on type, giving rise to overuse of specific individuals from a breed which had also recently undergone a genetic bottleneck.

Fell pony breeders have noted a familial pattern to FIS, with repeat matings producing diseased and normal offspring of both sexes. This, together with examination of the Fell and Dales stud books, which revealed small effective breeding populations and genetic bottlenecks, strongly suggested a genetic lesion, with an autosomal recessive mode of inheritance, as the cause of this disease. Autosomal recessive mutations are inherited in a basic Mendelian manner and are not sex linked, so affect both males and females. The sire and dam of affected offspring are asymptomatic carriers, having one copy of the disease allele and one copy of the wild-type allele. The offspring of carrier-carrier matings have a 25 per cent risk of inheriting two copies of the defective allele, giving rise to an affected animal (Table 1.1).

Table 1.1: *Expected outcomes (percentage of foals) from the matings of genotypically normal ponies and carriers of the FIS mutation.*

	Expected outcomes (%)		
	Normal foals	Carrier foals	Affected foals
Normal × Normal	100	0	0
Normal × Carrier	50	50	0
Carrier × Carrier	25	50	25

The generally low introduction of new breeding stock into these two breeds, and the genetic bottlenecks which they have undergone, would have undoubtedly impacted on the genetic variation within these two breeds. This in turn would have increased the relatedness, reducing the effective population size. The genetic lesion responsible for FIS would have arisen in a single animal. The transfer of this mutation to a popular stallion, which would have been used to cover multiple mares in any one breeding season in the Fell and Dales population, would enable the spread of this mutation into many more ponies. Over time, the number of carriers would have gradually increased and with a relatively small population, eventually consanguineous matings would give rise to affected offspring.

Comparable disease in the horse

There are no disorders in the horse that are known to cause both immunodeficiency and progressive anaemia. Based on this, it is likely that FIS is a novel disease. There are however several diseases with a genetic aetiology in the horse which cause either a deficiency in B-lymphocytes or a progressive anaemia.

Hereditary disorders of the horse which cause anaemia

Anaemia is defined as a reduction in the haemoglobin concentration of the blood, which is often accompanied by a reduced number of erythrocytes (Hoffbrand, 2001). Erythrocytes are derived from haemopoietic stem cells within the bone marrow; these are pluripotent stem cells which are progenitors of many cells, including lymphocytes and erythrocytes. Clinically, individuals with anaemia present with general signs, including pallor of mucous membranes, weakness and lethargy. There are three general classifications of anaemia; haemorrhagic, haemolytic and dyshaemopoietic. Examination of the anaemia suffered by FIS-affected foals reveals that it is unrelated to any blood loss or haemolysis and therefore dyshaemopoiesis is the most likely cause of the anaemia.

Two hereditary blood disorders have been reported in the horse: Von Willebrand disease in Thoroughbred and Quarter horses (Rathgeber et al., 2001, Laan et al., 2005), and Prekallikrein deficiency in Belgian and Miniature horses (Geor et al., 1990, Turrentine et al., 1986). However, neither of these conditions is likely to be the cause of the anaemia in FIS foals because both of these conditions are associated with haemorrhage.

Immunodeficiencies in the horse with known genetic aetiology

Immunodeficiency is generally defined as a condition in which the immune system becomes compromised, so is no longer able to fight infectious agents (Crisman and Scarratt, 2008). Both T and B-lymphocytes are derived from the haemopoietic stem

cells within the bone marrow. Then, under the influence of the thymus or secondary lymphoid organs, the progeny of the precursor cells further differentiate into T and B-lymphocytes. Normal equine foetuses are immunocompetent at birth (Perryman et al., 1980), with functional T-lymphocytes detected from 100 days gestation and B-lymphocytes from 200 days gestation. However, foals are immunologically naive at birth due to mares having an epitheliochorial placenta which does not allow the transfer of immunoglobulins in-utero. Therefore, ingestion of maternal antibodies from colostrum soon after birth is essential to provide the foal with some effective immunity against equine pathogens during the first few weeks of life.

Individuals with a compromised immune system have increased susceptibility to infectious agents, presenting with infections that require medical intervention. If the underlying immune dysfunction is not recognised, the horse will continue to suffer repeated infections, which can rapidly lead to failure to thrive and eventual death. Immunodeficiencies are characterised into two distinct groups – primary (genetic) and secondary (acquired) immunodeficiencies. Primary immunodeficiencies are relatively uncommon, arising from a genetic defect that determines immunological dysfunction, so the individual has increased susceptibility to infectious agents (Perryman, 2000). Due to the genetic nature of primary immunodeficiencies, clinically, they present a significant challenge to the veterinary practitioner, and despite persistent veterinary intervention, the individual will often die. Secondary immunodeficiencies do not have a genetic aetiology and can arise at any age from environmental and external contributing factors (Sellon, 2000).

Dysfunction of the immune system is usually categorised according to the specific components of the immunological system that are affected. Immunodeficiencies which involve only impairment of B-lymphocyte development and maturation are known as humoral immunodeficiencies (Buckley, 1986). Individuals may have absent or decreased expression of specific immunoglobulin classes. Clinically affected horses present with repeated infections from opportunistic organisms, which are persistent. Typical infections include *Streptococcus*, *Staphylococcus* and *Cryptosporidia*, presenting clinically as pneumonia, sinusitis, chronic diarrhoea and general failure to thrive. All of these infectious agents are only typically seen in individuals with a compromised immune system. Humoral immunodeficiencies

previously described in the horse include common variable immunodeficiency, selective immunoglobulin M (IgM) deficiency and agammaglobulinemia.

T-lymphocyte dysfunction, or B-lymphocyte dysfunction arising through lack of T-lymphocyte signalling, are known as cellular immune dysfunctions. Clinically, these are associated with infections of intracellular pathogens (viruses, protozoa, and mycobacteria) and are rarely observed in adult horses.

Combined immunodeficiencies include both B-lymphocyte (humoral) and T-lymphocyte (cellular) dysfunction and are fatal in horses (Crisman and Scarratt, 2008). Severe Combined Immunodeficiency (SCID) has been reported in multiple horse breeds, but is most commonly reported in the Arabian horse and is an autosomal recessive disease.

Pedigree analysis of FIS strongly suggests a genetic aetiology, with an autosomal recessive mode of inheritance. Therefore FIS can be classified as a primary immunodeficiency. Additionally, FIS has been characterised as a B-lymphocyte dysfunction with normal circulatory T-lymphocyte numbers, and based on this the immune dysfunction of FIS can be classified as a humoral immunodeficiency. Humoral immunodeficiencies which are known to have a genetic aetiology will now be reviewed, to consider if any of these disorders are a likely cause of FIS.

Selective Immunoglobulin M (IgM) Deficiency

Selective IgM deficiency has been reported in several breeds but is most commonly observed in Arabian and Quarter Horses (Perryman et al., 1977). This disorder has two clinical presentations. The first affects foals of 2 – 8 months of age and has a genetic aetiology. The second presentation is a secondary immunodeficiency, affecting horses older than two years and is associated with lymphosarcoma.

Foals with primary selective IgM deficiency are generally smaller and suffer repeated infections. Affected individuals respond well to antibiotic treatment but as soon as treatment ceases, the foal lapses and the infections become persistent. Most commonly, the respiratory tract is affected and foals often die or are euthanized

before eight months. Blood analysis reveals normal circulating levels of B-lymphocytes; some foals have a slightly reduced level of circulating T-lymphocytes and a slight anaemia. Immunoglobulin quantification of affected foals reveals a reduction in IgM concentration, which can be absolute in some cases, and normal levels of IgG, IgA and IgG(T) immunoglobulins (Weldon et al., 1992).

FIS-affected foals have a reduction in all immunoglobulin classes as well as a significant reduction in circulatory B-lymphocytes. In contrast, foals affected by selective immunoglobulin deficiency have normal circulatory B-lymphocyte numbers and normal levels of all immunoglobulin classes except IgM, which is significantly reduced. Based on the significant pathological differences between selective immunoglobulin deficiency and FIS, it is unlikely that selective immunoglobulin deficiency is the cause of FIS.

Agammaglobulinaemia

Agammaglobulinaemia has only been reported in male horses, all of which were either Thoroughbred, Standardbred, or Quarter horses (Crisman and Scarratt, 2008). It is likely, although not proven, that because all reported cases are male, agammaglobulinaemia is an X-linked disease. Affected individuals suffer repeated infections, with clinical signs including pneumonia, enteritis, dermatitis, arthritis and laminitis. This disorder is characterised by a complete absence of B-lymphocytes, IgM, IgA and very low serum concentrations of IgG and IgG(T) that decline as maternal antibodies decrease (Perryman et al., 1983). Foals present clinically between two and six months of age and initially respond well to antibiotic treatment, but when antibiotic treatment ceases repeated infections occur. All affected individuals die or are euthanized before two years of age.

Agammaglobulinaemia is similar to FIS in the respect that both conditions result in a severe depletion of B-lymphocytes. FIS-affected foals have a significant reduction in B-lymphocytes, which can be absolute in some cases, whereas in agammaglobulinaemia, all foals have absent B-lymphocytes. All immunoglobulin classes are significantly reduced in FIS-affected foals and similarly agammaglobulinaemia affected foals have a significant reduction in all immunoglobulin classes. However, clinically these conditions differ significantly in terms of expected life span. All FIS-affected foals die or are euthanized before 16

weeks due to infections becoming persistent and unresponsive to veterinary intervention, whereas foals with agammaglobulinaemia can survive until two years of age. Furthermore, FIS-affected foals have a progressive anaemia, which has not been reported in any agammaglobulinaemia cases. Based on the life expectancy of agammaglobulinaemia affected individuals compared to FIS-affected foals, it is highly unlikely that agammaglobulinaemia is the cause of FIS.

Both of these conditions have some similarities to FIS. However, neither describes one of the primary characteristics of FIS, a severe progressive anaemia, which is observed in all FIS-affected individuals. Given the specific combination of anaemia, B-lymphocyte deficiency and reduction in all immunoglobulin classes, it is highly probable that FIS is a novel equine disorder.

Human genetic disorders that cause immunodeficiency and anaemia

Comparable diseases may differ in clinical presentation between alternative species. Therefore, genetic disorders in man, which present as an immunodeficiency with a variable anaemia, will be explored as the cause of FIS.

Shwachman-Diamond syndrome (SDS), first described in the 1960's, is an autosomal recessive disorder that has a multisystem effect. Approximately 90% of reported cases have a mutation in the *SBDS* gene (Shwachman-Bodian-Diamond syndrome gene), which is highly expressed in multiple human tissues at both the messenger RNA (mRNA) and protein levels. The *SBDS* gene is located on chromosome seven in the human and is predicted to be on chromosome 13 in the equine genome. Primary characteristics of the disorder include bone-marrow failure, pancreatic dysfunction and skeletal abnormalities. Anaemia has been reported in up to 80% of affected individuals and is usually normochromic–normocytic (Dror, 2005). Individuals with SDS are highly susceptible to recurrent viral, bacterial and fungal infections, which result in premature death, at a young age. Common infections associated with SDS include sinusitis, mouth sores, bronchopneumonia, septicaemia, and infections of the skin. Both B and T-lymphocyte defects have been reported, specifically decreased circulatory B-lymphocytes, low immunoglobulin levels, decreased percentages of circulating natural killer (NK) cells and reduced total circulating T lymphocytes (Burroughs et al., 2009). Individuals with SDS also present with abnormal skeletal development, malabsorption, failure to thrive, and low levels of fat soluble vitamins A, D, E, and K which result from pancreatic dysfunction. Affected individuals can be clinically managed, surviving into adulthood. Most commonly, premature death results from malabsorption and infections, and in older patients, death occurs from haemorrhage and infections due to associated haematological abnormalities.

SDS is similar to FIS in its clinical presentation, as both syndromes result in immunodeficiency and anaemia. However skeletal abnormalities and a T-

lymphocyte deficiency, both primary characteristics of SDS, have not been observed in FIS. It is therefore unlikely that SDS is the cause of FIS.

Cartilage-hair hypoplasia (CHH) is an autosomal recessive disorder, which is associated with short limbed short stature, hypoplastic hair, variable anaemia and a mild to moderately severe cellular immunodeficiency (Makitie et al., 1995). CHH is associated with mutations of the *RMRP* gene, although the exact pathogenesis of the disease remains unknown. The *RMRP* gene is located on human chromosome nine, with genomic alignments predicting that this gene is on chromosome 25 in the horse. A large number of CHH affected individuals have a T-lymphocyte deficiency, with decreased and delayed responsiveness of both B and T-lymphocytes (Polmar and Pierce, 1986). Usually CHH individuals present in early childhood with mild macrocytic anaemia, due to defective erythrocyte production. However, erythrocyte production is usually normal by adulthood, spontaneously correcting itself. The primary characteristic of CHH is dwarfism, which is associated with all CHH affected individuals, a characteristic that makes CHH an unlikely cause of FIS.

Both of these human syndromes have similarities with FIS as both comprise a B-lymphocyte deficiency and a variable anaemia. However, FIS is a lethal condition, with all affected individuals dying before 16 weeks of age, whereas SDS and CHH affected individuals can be clinically managed to survive into adulthood. Therefore, given the specific characteristics of FIS and how they differ from SDS and CHH, it is likely that FIS is a novel disease which has does not have a documented parallel in man.

Summary of Foal Immunodeficiency Syndrome

FIS is a novel disease that results in a progressive profound anaemia and a severe immunodeficiency, comprising B-lymphocyte deficiency and a severe reduction in all immunoglobulin classes. Both of the affected cell types are derived from the haemopoietic stem cell in the bone marrow, which suggests that the bone marrow is severely affected by this disease. Breeders have noted a familial pattern with both sexes affected. Parents are apparently completely normal, which suggests that FIS has a genetic aetiology, which is autosomal recessive. Pedigree analysis suggests bottlenecks in both the Fell and Dales, with a popular sire effect, resulting in a loss of genetic diversity in the Fell Pony.

Genetic investigation, to identify the causal variant, is critical for the survival of both the Fell and Dales pony, which are both registered by the Rare Breeds Survival Trust. Identification of the causal variant would ultimately lead to the introduction of a genetic test. This could then be used to screen pony breeding stocks for carriers of the mutation, and with careful breeding programs, rid these populations of this lethal disease. To date, research has focused on characterising and understanding the pathological effects of this disease rather than investigating the genetics, most probably due to the limited tools available. However in January 2007, the equine genome sequence became publically available, which has in-turn provided new tools for disease mapping which will enable efficient and successful mapping of FIS.

Aims and Objectives of this investigation

i. Aims

1. The primary aim of this project is to characterise the genetic defect which leads to the disease known as Foal Immunodeficiency Syndrome
2. Develop a diagnostic test that can be used in practice to screen Fell and Dales Pony breeding stocks for FIS-carriers. Further to this, the test will be used as a diagnostic aid, to diagnose FIS-affected foals, and to determine if the mutation has spread into other horse breeds.

3. A secondary aim of this project is to identify downstream genes whose activity is altered by FPS, and correlate this with the disease phenotype as far as possible.

ii. Objectives

1. Using genome-wide mapping approaches, identify the chromosome where the FIS mutation resides.
2. Fine map the chromosomal region that shows significant association with the FIS mutation, using microsatellites and SNPs. This would lead to the identification of the homozygous identical-by-descent haplotype on which the mutation arose and could be the location of the FIS mutation.
3. Identify candidate genes within the critical region, based on function and phenotypic similarities with other species. These plausible candidate genes will then be sequenced in an attempt to identify the causal mutation.
4. Should the causal mutation not be identified from sequencing candidate genes, alternative high-throughput re-sequencing methods will be adopted to interrogate the entire critical region, which will lead to the identification of the causal variant.
5. Examine the gene in which the causal variant is identified, to provide an understanding of how this gene could potentially cause FIS.
6. Perform a population screen of the Fell and Dales pony, to estimate the prevalence of FIS-carriers in the general population.
7. Identify additional breeds that may be at-risk from the transfer of the FIS mutation via interbreeding with the Fell and Dales. Perform a population screen of these breeds to assess the spread of the FIS mutation into these populations.
8. Collect bone marrow (FIS and healthy age matched control) samples for gene expression analysis. Perform global transcriptional analysis to identify those genes which are most significantly up- or down-regulated by the disease. Further to this, networking analysis will be performed to examine those pathways which are most severely affected.

Investigational summary

Chapter two describes how microsatellite markers were used in a whole-genome scan to identify a chromosomal location which displays significant linkage to FIS. In Chapter three, the results from a genome-wide association study using single nucleotide polymorphisms, are presented. This provided further evidence that the chromosomal location identified from microsatellite mapping, was the most likely location of the FIS mutation. Identification of the chromosomal location of the FIS mutation provided the basis for fine-mapping which is discussed in chapter four, identifying the homozygous identical-by-descent (IBD) haplotype on which the FIS mutation arose. Chapter five examines the homozygous haplotype using next-generation re-sequencing, which lead to the identification of a highly associated mutation, which is then utilised in Chapter six to perform a population study. In Chapter seven, a pilot study is performed to investigate global transcriptional changes in FIS-affected foals compared to healthy controls. Finally, Chapter eight discusses the implications of the findings of the proceeding chapters, discussing the identified mutation and how it may result in the FIS phenotype. This chapter also highlights future work which may provide further insight into the genetics of FIS.

Chapter 2

Mapping an Associated Disease Locus Using Microsatellite Markers

	Page
Summary	33
2.1 Introduction	33
Disease mapping with microsatellite markers	33
Basic principles of genotyping microsatellites	34
Mapping traits using linkage mapping	35
Homozygosity mapping	37
Objectives and aims of the FIS whole-genome microsatellite mapping approach	39
2.2 Materials and Methods	40
Animals and Samples	40
DNA extraction from tissue and blood samples	41
Calculation of power to detect significant linkage	41
Microsatellite genome scan	41
Statistical analysis of the genotyping data	43
2.3 Results	44
Two-point linkage analysis results	47
Results of the homozygosity mapping	47
Multi-point linkage analysis	49
2.4 Discussion	52

Summary

Stud book analysis, incorporating the identity of affected animals, strongly suggested that a genetic lesion, which is autosomal recessive in nature, is responsible for FIS (described in Chapter 1). Both a homozygosity mapping and a linkage based mapping approach were used in attempt to map the associated locus. A whole-genome microsatellite marker scan was performed to assess linkage and to search for regions of homozygosity. This led to the identification of one chromosome which was significantly associated with FIS.

2.1 Introduction

Disease mapping with microsatellite markers

Microsatellite markers are loci of heterogeneity within the genome and have been used to successfully map many heritable traits in the horse (Cook et al., 2008, Andersson et al., 2008, Terry et al., 2004). Microsatellite markers are short variable number tandem repeats (VNTRs) of a simple sequence, typically 2-4bp, most commonly the dinucleotide (CA)_n and its complement (GT)_n. Microsatellites are very polymorphic and so highly informative for disease mapping. They have the added advantage that multiple markers can be amplified together, increasing efficiency and reducing overall costs. The first equine microsatellite markers were published nearly 20 years ago (Ellegren et al., 1992) and since then, much effort has been devoted to identifying markers and developing comprehensive equine linkage maps, to enable further progress in mapping traits. The first horse linkage map was produced in 2000 (Swinburne et al., 2000) using two three-generation, full-sibling, crossbred horse reference families. Three years after this, the second generation linkage map was published (Guerin et al., 2003), followed by another more comprehensive horse linkage map in 2006 (Swinburne et al., 2006). Additionally, further microsatellites for linkage and RH maps have also been developed (Tozaki et al., 2007). Currently (query performed September 2010), 24,109 microsatellite sequences have been submitted to the National Centre for Biotechnology Information (NCBI) database and are freely available on the internet

(<http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucore&cmd=search&term=equine%20microsatellite>).

Basic principles of genotyping microsatellites

Isolating DNA for genetic investigation

Genomic DNA (gDNA) isolated from cells is used for genetic investigation. A range of material can be used for the isolation of gDNA, including blood, tissue, hair and saliva. Irrespective of the source, extraction and purification of DNA involves complete disruption and lysis of cells walls and plasma membranes of cells and organelles, to yield a solution of DNA, RNA and other non-nucleic acid components (e.g. protein). The purified DNA is then isolated using either filters which bind the DNA which is then washed and eluted, or through phase separation by centrifugation followed by ethanol precipitation and re-suspension in an appropriate volume of buffer. The purified DNA is then quantified and diluted to the appropriate concentration for the Polymerase Chain Reaction (PCR) which will amplify the target sequence.

Multiplexing microsatellite markers for increased efficiency

Much effort has been devoted to producing compatible microsatellite markers that can be co-amplified in a single multiplex PCR reaction (Swinburne et al., 2002), increasing efficiency and reducing costs. Markers selected for the whole-genome scan are divided into panels, chosen so that they have non-overlapping allele sizes so they can be analysed in the same run. The panels would then be further sub-divided into groups that are compatible for PCR amplification. The PCR products are then labelled during amplification, using a method such as fluorescent labelling (figure 2.1), which enables post-amplification pooling of multiple amplicons with non-overlapping allele sizes for analysis using capillary electrophoresis. The genotypes are then visualised using genotyping software so the appropriate allele can be called ready for statistical analysis. After completing allele assignment, the data can then be analysed to identify a locus displaying significant disease linkage, using a linkage based or homozygosity mapping approach.

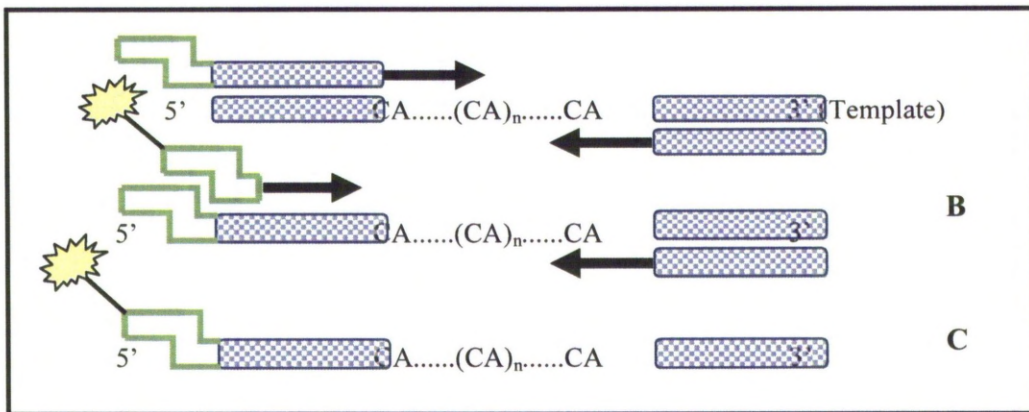


Figure 2.1: PCR amplicon labelling using the 3-primer fluorescent methodology (adapted from (Schuelke, 2000)). **A:** In the first PCR cycles, the sequence specific forward primer with a primer tail is incorporated into the PCR products. **B:** These PCR products with a tail are the target for the fluorescently labelled primer, which is incorporated during subsequent cycles at a lower annealing temperature. **C:** The fluorescently labelled final product can then be analysed using a laser detection system.

Mapping traits using linkage mapping approach

Linkage mapping involves typing microsatellite markers across the genome within a family (over a few generations) that are affected by the trait of interest. Using this data, analysis is performed in an attempt to identify a chromosomal region which is inherited with the trait more often than would be expected due to chance (Hirschhorn and Daly, 2005, Dawn Teare and Barrett, 2005). The identified region in which the marker and disease variant have been co-inherited is likely to contain the causal genetic variant (Fig. 2.2).

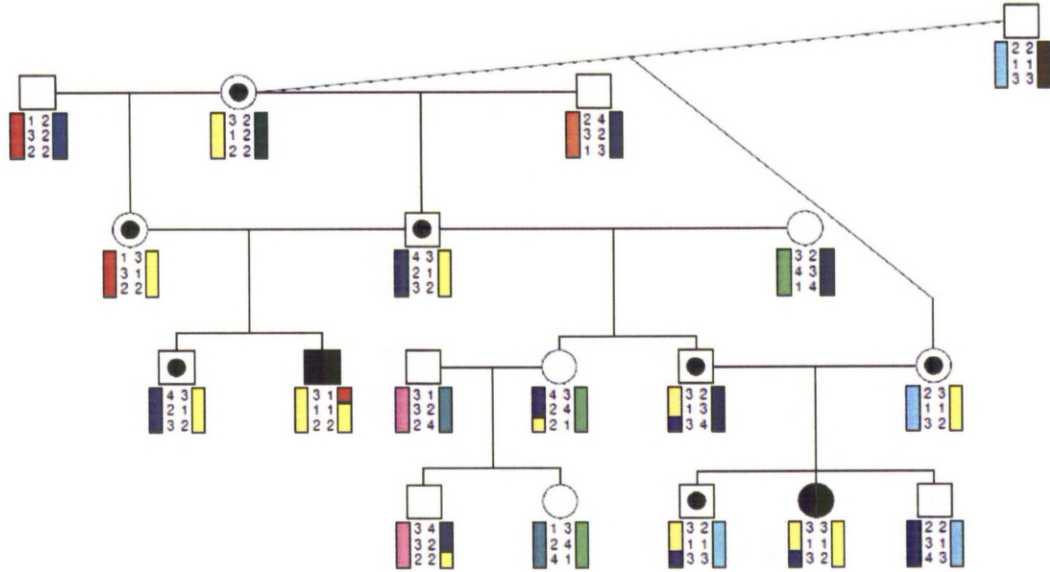


Figure 2.2: The principle of linkage analysis for an autosomal recessive disease in a family which has been typed for three markers. The pedigree contains linkage information showing that the yellow haplotype shows linkage and marker 1 on the yellow haplotype is co-segregating with the disease. Using linkage software, the overall logarithm of odds (LOD score) can be calculated, on the alternative assumptions that the loci are linked (recombination fraction = θ) or not linked (recombination fraction = 0.5). Numbers represent genotypic markers and colours represent haplotypes to enable the tracing of haplotype inheritance. Squares indicate males and circles females. Those with a dot are obligate carriers and filled circles/squares are affected by the disease in question.

Linkage based analysis (parametric analysis) is used to assess the degree of co-segregation of alleles within a pedigree, searching for homozygous IBD chromosomal segments which are shared by all of the affected samples. The standard format for reporting linkage is as a logarithm of the odds (LOD) score which represents the most efficient statistic for evaluating pedigree linkage (Morton, 1955). The higher the calculated LOD score the greater the disease-marker association, with LOD scores of >3 being considered significant evidence of linkage. Linkage analysis can either be performed as two-point analysis or as multipoint analysis. Two-point mapping is usually used as the primary analysis, assigning the disease association to an individual chromosome. Each marker is analysed independently, calculating a LOD score for the overall likelihood that the marker and disease locus are linked. In

contrast, multipoint linkage analysis considers the disease trait in combination with multiple loci, enabling localisation of the disease association between two markers. The program works by moving the disease gene around the specified marker framework, calculating the overall likelihood (LOD score) of the pedigree data at each position.

Whole-genome linkage based studies have successfully mapped many monogenic ‘Mendelian’ traits but when considering a study of this type, a key question is how many markers are required to provide sufficient coverage of the genome. This is because in order for the mapping to be successful, there must be adequate markers spaced across the genome to enable the detection of a marker which is co-segregating with the disease allele. Successful trait mapping has been seen using as few as 41 microsatellite markers in the horse (Brunberg et al., 2006), although a recent study estimated that ~322 microsatellite markers are required to detect reliable disease linkage in the horse (Mittmann et al., 2010). A further consideration with linkage based studies is that large sample numbers are required within the pedigree to provide sufficient power to detect statistically significant disease linkage. An approach which has been developed to overcome the limitations of sample collection within families is the homozygosity mapping approach.

Homozygosity mapping

Homozygosity mapping, first described in 1987 (Lander and Botstein, 1987) is an approach which has been used to successfully map rare Mendelian traits using minimal numbers of affected cases (Escamilla et al., 2000). This approach does not require pedigree information. Therefore this approach lends itself as both an independent approach or as complementary analysis to linkage based studies when sample numbers limit the power to detect significant linkage. Homozygosity mapping utilises the fact that offspring from consanguineous matings that are affected with the same recessive genetic disease, will share a region of homozygosity. The IBD homozygous region will span the disease locus, representing the ancestral haplotype on which the mutation arose, pointing the investigator to the chromosomal region that harbours the causal mutation. Multiple regions of homozygosity may be detected in any given offspring, but looking for

regions that are consistently homozygous in multiple samples provides a powerful strategy for mapping a recessive gene (Lander and Botstein, 1987).

As homozygosity mapping requires no pedigree information or assumptions about the population to be tested, non-parametric analyses are used to identify the shared IBD homozygous region. The associated regions are identified by simply detecting a significant difference in allele frequency between the affected and unaffected groups; this can be performed using statistical methods which compare groups, such as Chi^2 analysis, or by using computed algorithms that identify a loss of heterozygosity.

Objectives and aims of the FIS whole-genome microsatellite mapping approach

Due to the published microsatellite sequences and the success of mapping Mendelian traits in the horse using whole-genome microsatellite scans (Terry et al., 2004, Tryon et al., 2007, Swinburne et al., 2002), this approach was used in attempt to map an associated FIS disease locus. The microsatellite genotyping data was analysed using both the parametric linkage based approach and the non-parametric homozygosity analysis. The non-parametric homozygosity analysis was used to supplement the findings of the linkage based study, accounting for the relatively few samples that have been used in this study.

The aims of this microsatellite genome scan were to:

- a) Identify markers showing significant linkage to the FIS phenotype to identify a critical chromosome region for further analysis.
- b) Identify regions which were significantly more homozygous in the affected animals compared to the unaffected using Chi squared analysis.

2.2 Materials and Methods

Animals and Samples

Samples used for the microsatellite genotyping were collected as part of Dr. Gareth Thomas's PhD project, titled 'Immunodeficiency in Fell Ponies' (Thomas, 2003). All of these samples were from Fell Ponies. Samples were collected from affected foals, parents of affected foals (obligate carriers), individuals of unknown carrier status and adult samples with unknown carrier status. All samples were collected under the Veterinary Surgeons Act 1966, inflicting minimal pain on living animals. Disease status of suspected FIS foals was confirmed post-mortem based on haematological and histological examination and gross findings (Scholes et al., 1998). In addition, whenever possible, bone marrow samples collected by Dr. G Thomas were also assessed for reduced numbers of erythrocytes (Thomas, 2003).

Jugular blood samples were collected by Dr. G Thomas or by a veterinary surgeon in private practice and posted to the University of Liverpool for haematological profiling; excess was used for DNA extractions. Samples were collected into Vacutainers® containing anticoagulant (Beckton Dickinson, UK) before archiving at -20°C. Only samples collected into ethylenediamine tetra-acetic acid (EDTA) were used in this study.

Tissue samples (kidney, spleen, muscle and lymph node) were collected during post-mortem examination from some of the affected foals by the Veterinary Investigation Centre (Merrythought), Penrith and the Department of Veterinary Pathology, University of Liverpool. These were archived at -20°C.

Samples were identified from 44 (appendix 1) animals in five small pedigrees for genetic analysis. Two of these were tissue samples from affected foals and the remaining 42 samples were blood samples. The sample set consisted of 18 obligate carriers (parents to confirmed affected foals), 14 affected foals and 12 adult samples of unknown carrier status. Further pedigree analysis enabled the joining of these 5 pedigrees into a single pedigree, which was used for the linkage analysis.

DNA extraction from tissue and blood samples

DNA extractions for the samples used in the microsatellite genome scan were performed by Ms. G Hill using the Nucleon® genomic DNA (gDNA) extraction kit, a commercially available kit available from GE Healthcare, UK. DNA was extracted using a revised version of the manufacturer's instructions (Appendix 2) and re-suspended in ddH₂O. The concentration of the DNA sample was assessed using a Nanodrop®, by measuring absorbance at 260nm, according to the manufacturer's instructions. An aliquot of stock DNA was then further diluted to an approximate concentration of 10ng/μL using ddH₂O.

Calculation of power to detect significant linkage

An estimation of the maximum LOD score expected when performing linkage analysis with these samples was generated by performing simulations using SLINK (<http://linkage.rockefeller.edu/ott/SLINK.htm>). The simulations were performed assuming a disease allele frequency of 0.14, 100% penetrance and a simulated marker with two alleles each at a frequency of 0.5.

Microsatellite genome scan

A panel of 286 microsatellite markers, were developed at the Animal Health Trust and previously used to successfully map disease loci (Swinburne et al., 2009), was used in this study. The marker set was developed with reference to the horse linkage map (Swinburne et al., 2006) to provide a comprehensive set for performing a low-density scan of all 31 horse autosomes and the X chromosome (Appendix 3 lists all of the markers used in this study). In addition, five microsatellite markers (TKY766, TKY502, COR071, COR099 and TKY1155) were genotyped individually. The markers used in this study were selected with reference to the equine linkage map, the physical map not yet being available. The subsequent sequencing of the equine genome enabled the physical position of the markers to be later confirmed. The physical position of all markers was confirmed on the horse genome assembly

EquCab2.0 using BLAST-like alignment tool (BLAT) in Ensembl (<http://www.ensembl.org/Multi/blastview>). Markers where multiple alignments were observed and the position was not conclusive were excluded from further analysis.

The genome scan was performed in polymerase chain reaction (PCR) multiplexes of three markers per reaction. Four PCR reactions, each utilising a different fluorescent dye, were pooled together post-PCR to form a panel of 12 markers for analysis. Fluorescent labelling of the PCR amplicon was achieved using the 3-primer methodology. All aliquoting, PCR set up, and pooling steps were performed using a Thermo Scientific Matrix PlateMate 2×2 automated pipetting workstation (Thermo Fisher Scientific, Waltham, MA). Amplification for fragment analysis was performed in 384-well PCR plates (Axygen Scientific, Union City, CA) in 6µl volumes, using 20ng gDNA, 0.75 unit AmpliTaq Gold (Applied Biosystems, Foster City, CA), 1 × GeneAmp PCR buffer II (Applied Biosystems), 1.5mM MgCl₂, and 200µM each dNTP. Then 2.5 pmol of reverse, 1 pmol of tailed-forward, and 5 pmol of the labelled universal primer (either 6-FAM, VIC, NED, or PET) were added to the reaction. Samples were run on an MJ Tetrad PCR cycler (Bio-Rad Laboratories, Hercules, CA). A PCR program of 94°C for 10 min, followed by 30 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by 8 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 1 min, and then 72°C for 30 min was used. A 2µl aliquot of the four PCR reactions belonging to each panel of 12 markers was then pooled together for analysis. Reactions were stored at -20°C prior to loading a 3µl aliquot for analysis on an ABI3100 (Applied Biosystems) according to the manufacturer's instructions. Dye set G5 was used in conjunction with the LIZ500 size standard.

Genotyping data was collected and analysed with GeneMapper version 4.0 (Applied Biosystems). Allele sizes were assigned to pre-defined bins and automatically given an appropriate integer value. Genotypes were also scored manually to check for errors. Data was then exported into Excel and assessed for monomorphic markers, which were excluded from further analysis.

Statistical analysis of the genotyping data

Initially, allele and genotype frequencies were counted within the control population and the affected population for polymorphic autosomal microsatellite loci. All subsequent analysis was only performed on the autosomal markers.

Pedigree data and genotypes were imported into Progeny (Progeny Software LLC, South Bend, IN, www.progenygenetics.com) and pedigrees constructed. The data was then assessed for inconsistent inheritance (non-Mendelian inheritance when parental genotypes were available). The FIS mutation was defined as fully penetrant, autosomal and recessive.

Linkage based analysis

Two-point mapping was used to identify associated markers, using the online version of SUPERLINK v1.7 (<http://bioinfo.cs.technion.ac.il/superlink/>) (Silberstein et al., 2006). The single locus disease model was used with a disease frequency of 0.1. In order to further fine-map any associations, a multipoint analysis was then performed. Again, the single disease model was selected with a disease frequency of 0.1. LOD-scores were calculated at every marker and at two points between adjacent markers. LOD-scores were also calculated 10cM before the first marker and 10cM after the last marker on each chromosome.

Homozygosity mapping

As a complementary analysis the Pearson's χ^2 test of independence was used to test the null hypothesis that the two populations (control and affected) have the same allele frequencies. An $A \times 2$ (where A = No. of alleles present at a given locus) contingency table with $A-1$ degrees of freedom were used. In addition expected and observed heterozygosity values were computed for markers exhibiting a positive LOD score, using Arlequin version 3.5 (<http://cmpg.unibe.ch/software/arlequin35/>) (Excoffier et al., 2005).

2.3 Results

Using the BLAST-like alignment tool (BLAT) in Ensembl (<http://www.ensembl.org/Multi/blastview>), the physical position of the microsatellite markers were identified on the EquCab2.0 build of the equine genome. All markers where multiple alignments were observed and the position was not conclusive were excluded from further analysis. Of the 286 multiplex microsatellite markers genotyped, 58 were excluded from further analysis. The discarded markers included 10 *Equus caballus* (ECA) X markers (because the inheritance of FIS is not sex-linked); 37 markers that were either monomorphic ($n = 2$) or did not genotype well ($n = 34$); nine markers that could not be conclusively positioned on the physical map; and two markers which mapped to the same locus as another so were presumed to be duplicates. The five individual additional markers were all included in further analysis. The average spacing of the 233 markers on the physical map which were used for analysis was 9.01 Mb, with a range of 0.5 - 42.97 Mb. A complete list of all the markers used for analysis can be found in Appendix 4.

Of the 44 samples genotyped, 3 animals were excluded as they either did not genotype well ($n = 2$) or were identified as having inconsistent inheritance based on pedigree information ($n = 1$). Therefore, 41 samples (Fig. 2.3) were used for the linkage analysis of FIS, consisting of 13 affected FIS animals, 17 obligate carrier samples and 11 adult pony samples of unknown carrier samples. Further pedigree analysis enabled the joining of these 41 samples into one single pedigree (Fig. 2.4).

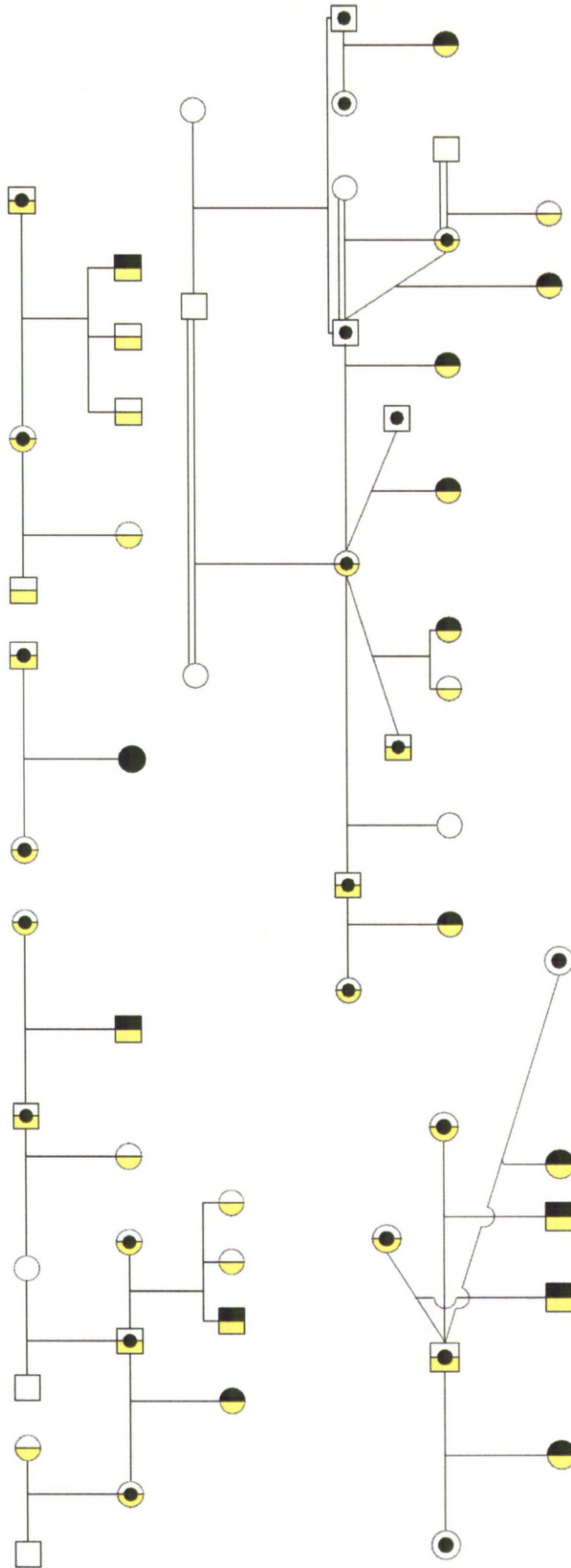


Figure 2.3: Five small pedigrees which show the relationships of the forty-one related animals which were selected and used for genetic analysis with microsatellite markers. FIS-affected individuals are shown shaded in black and obligate carriers are indicated with a dot. Individuals which are not affected or obligate carriers are shown un-shaded. Individuals also coloured yellow were genotyped and used for linkage and homozygosity mapping. Double lines indicate consanguinity.

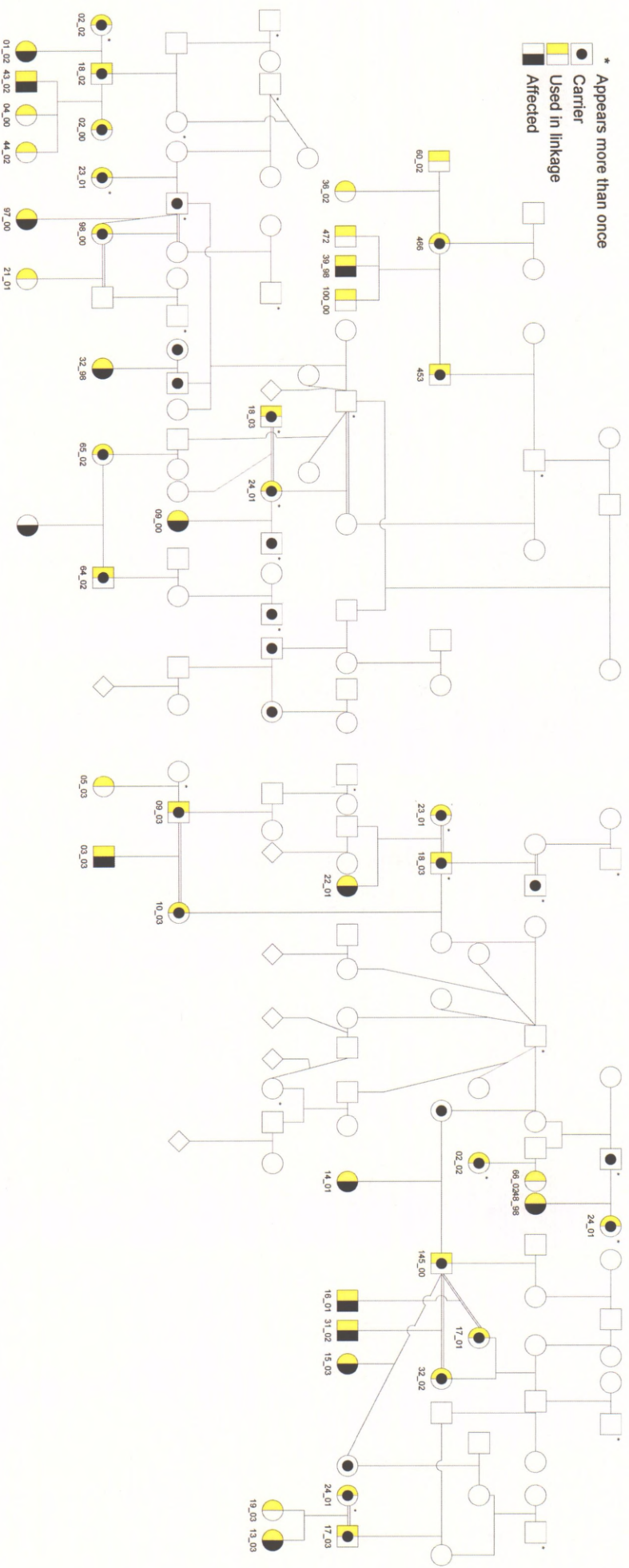


Figure 2.4: Extended pedigree structure which was constructed in Progeny, incorporating the five small pedigrees shown in Fig. 2.3 and used in linkage analysis for mapping the FIS associated loci. Those samples with a star are shadowed samples (a replicate of a sample), which was required to break pedigree loops for linkage analysis.

Estimation of the maximum expected LOD score

A simulated marker was used to calculate the maximum expected LOD score which could be expected from the linkage analysis. Simulations revealed that a maximum LOD score of 3.6350 would be observed for a marker showing disease linkage at a recombination fraction of 0 (Table 2.1).

LOD SCORE AT VARIOUS RECOMBINATION FRACTIONS BETWEEN SIMULATED MARKER AND MUTATION						
0	0.01	0.05	0.1	0.2	0.3	0.4
3.6350	3.5209	3.0706	2.5264	1.5343	0.7310	0.1926

Table 2.1. *Estimations of the maximum LOD score expected at various recombination fractions between disease and microsatellite marker.*

Two-point linkage analysis results

Of the 233 microsatellite loci tested, two-point mapping revealed 21 markers exhibiting a LOD score >1 on 16 chromosomes, of which only 1 was statistically significant (NVHEQ070 on ECA26) with a LOD score exceeding 3 (table 2.2).

Results of the homozygosity mapping

Potentially associated loci were further evaluated for significant differences in allele frequencies using the Pearson's Chi² test of independence by comparing the observed heterozygosities in the affected population (n=13) with the control population (n=28). Only one marker, NVHEQ070 gave a significant P-value (Table 2.2). Heterozygosity values were computed for the affected and unaffected groups using Arlequin.

Marker Name	Chrom.	Position (Mb)	LOD SCORE AT VARIOUS RECOMBINATION FRACTIONS										Observed Heterozygosity		Chi squared test of independence	
			0	0.01	0.05	0.1	0.2	0.3	0.4	Control	Affected	Chi Squared	d.f.	p-value		
HLM005	1	1.63	1.9075	1.873	1.7104	1.4764	0.9861	0.545	0.2101	0.4615	0.1250	1.177	1	0.2779		
SGCV002	1	76.35	1.0005	0.9581	0.7974	0.6193	0.3469	0.1738	0.0681	0.1539	0.1250	0.000	1	1.0000		
VHL123A	2	109.82	2.9745	2.8954	2.5803	2.1929	1.4609	0.8206	0.3171	0.1539	0.0625	2.518	2	0.2840		
TKY223	4	8.65	-5.3045	-0.3098	0.7476	1.0757	0.9505	0.5656	0.2267	0.65385	0.55847	3.785	4	0.4360		
LEX014	5	88.96	-2.3497	-0.1788	0.7841	1.0441	0.9766	0.6774	0.3262	0.74133	0.71371	1.334	3	0.7210		
UM237	5	97.46	0.3941	0.8088	1.1794	1.1107	0.644	0.2227	0.0169	0.70211	0.74395	1.782	3	0.6190		
TKY312	6	17.32	0.6735	0.7829	0.9916	1.0179	0.8082	0.5069	0.2242	0.55581	0.66734	2.029	3	0.5664		
TKY005	7	43.60	1.2685	1.2517	1.1574	1.0079	0.6868	0.3961	0.1651	0.1923	0.1875	0.277	2	0.8705		
TKY131	10	10.00	0.2606	0.4734	0.94	1.1282	1.0069	0.6424	0.2653	0.6923	0.4375	6.265	6	0.3942		
UCDEQ497	12	32.57	1.3798	1.3604	1.2531	1.0832	0.7201	0.4032	0.1652	0.1539	0.1250	1.774	2	0.4118		
UM010	14	25.47	-0.1678	0.8671	1.4252	1.461	1.0801	0.5786	0.1938	0.6923	0.3750	3.708	6	0.7161		
TKY491	14	81.18	-1.2065	0.2821	1.2125	1.3198	0.9654	0.5166	0.1886	0.5769	0.5625	1.023	2	0.5997		
HMS001	15	85.45	1.3257	1.3119	1.2419	1.0833	0.6604	0.3128	0.1089	0.1539	0.4375	3.782	5	0.5812		
AHT014	16	57.79	1.2171	1.1842	1.0532	0.8925	0.5905	0.3297	0.1296	0.1923	0.0000	2.644	2	0.2666		
HMS058	16	81.91	0.1724	0.7087	1.3322	1.4534	1.1536	0.6782	0.265	0.3846	0.3750	2.310	4	0.6789		
UM022	23	32.97	2.5894	2.5528	2.3325	1.9806	1.242	0.6201	0.202	0.5000	0.2500	5.454	4	0.2438		
TKY394	24	33.98	-0.5672	1.5628	2.1335	2.0723	1.5078	0.8699	0.3501	0.4615	0.3125	3.227	3	0.3580		
TKY1155	26	29.81	1.6842	1.6977	1.6447	1.4654	0.9973	0.5374	0.1785	0.5769	0.3750	0.166	3	0.9828		
NVHEQ070	26	30.25	3.2914	3.1966	2.8239	2.3759	1.5567	0.8631	0.3289	0.4615	0.0625	7.150	2	0.0280		
AHT082	27	27.27	1.808	1.908	1.9326	1.7373	1.1975	0.6582	0.229	0.2692	0.3125	5.360	3	0.1473		
UMNE530	30	11.73	1.3547	1.4223	1.4427	1.2852	0.8494	0.4514	0.1653	0.4615	0.1250	3.032	3	0.3867		

Table 2.2: Two-point linkage results for markers which showed LOD score of greater than one; a LOD score of 3.2914 was obtained with microsatellite marker NVHEQ070, which also showed a significant difference in allele frequency between affecteds and controls. Significant LOD scores are shown in orange (greater than three), LOD scores less than three but greater than two are shown in pale orange, LOD scores less than two but greater than one are shown in yellow, those in which were less than one but greater than zero are shown in white and those that were less than zero are indicated in blue. Significant p-values for the Pearson's Chi squared test of independence are shown in orange.

Multi-point linkage analysis results

To further assess disease linkage multi-point LOD scores were computed for chromosomes where a LOD score >1 had been observed with the two-point mapping. Only two chromosomes exhibited a LOD score >1 (ECA6 and ECA26), with the maximum LOD score of 3.42 generated at marker NVHEQ070 on ECA26 (Fig. 2.5). With the exception of flanking markers within 3.73 Mb of NVHEQ070, no other marker in the genome generated a LOD score >3.0 . Two separate regions on ECA6 showed disease linkage, with the maximum LOD score of 2.07 observed at marker TKY360 (Fig. 2.5). This region spans 35.7 Mb, encapsulating 3 microsatellite markers, including HMS055 which also showed displayed linkage with two-point mapping. The second lesser association was observed at marker UM177, with a maximum LOD score of 1.62, spanning a 14 Mb region (10 Mb of which is off the end of the marker map). A total of 9 chromosomes exhibited LOD scores greater than zero (Table 2.3), with an average LOD score of -2.793 (range -0.0209 to -13.73) observed for all remaining 20 chromosomes.

Chrom.	Position	Marker	LOD Score
1	72.95		0.3158
1	74.65		0.5595
1	76.35	SGCV002	0.682
1	80.85		0.7634
1	85.35		0.7125
1	89.89	AHT040	0.5075
1	90.46		0.3211
1	91.03		0.01
1	99.98	ASB008	0.0773
1	103.42		0.041
1	182.81	COR053	0.0638
1	186.15		0.061
1	189.48		0.0533
1	192.81		0.0437
2	109.82	VHL123A	0.2912
2	112.25		0.4048
2	114.72		0.4962
2	117.18	COR026	0.5719
2	120.52		0.5682
2	123.85		0.5546
2	127.18		0.5347
6	1.26		1.4995
6	4.59		1.5752
6	7.93		1.6262
6	11.26	UM177	1.6196
6	13.29		1.3883
6	15.33		1.0654
6	17.32	TKY312	0.5092
6	18.22		0.3898
6	19.12		0.1556
6	24.58		1.0064
6	29.18		1.4979
6	33.74	HMS055	1.8187
6	34.51		1.8562
6	35.27		1.8923
6	36.03	TKY377	1.9273
6	38.10		1.9909
6	40.17		2.0370
6	42.22	TKY360	2.0706
6	51.22		1.9280
6	60.22		1.6730
6	74.39		0.8685
6	79.56		0.7233
6	88.05		0.0135
6	91.39		0.2034
6	94.72		0.2960

Chrom.	Position	Marker	LOD Score
7	38.73		0.0691
7	43.60	TKY005	0.3092
7	43.60		0.3184
7	50.66		0.1382
10	5.13		0.0008
10	7.56		0.3814
10	10.00	TKY131	0.0561
10	10.93		0.5915
10	11.86		0.5555
11	54.82		0.053
11	58.15		0.1411
26	1.34		0.6767
26	4.67		0.6506
26	8.00		0.5697
26	11.34	TKY766	0.3418
26	12.27		0.0746
26	23.91		0.1235
26	26.04	UM005	0.2142
26	28.29		1.9047
26	29.06		2.8137
26	29.81	TKY1155	3.3587
26	29.94		3.3812
26	30.07		3.4025
26	30.25	NVHEQ070	3.4224
26	33.59		3.1585
26	36.92		2.9100
26	40.25		2.6769
27	38.36		0.2747
27	41.70		0.3847
29	8.88		0.2459
30	10.26		0.1986
30	11.73	UMNE530	0.5504
30	16.96		0.6385
30	22.20		0.7056
30	27.41	UCDEQ455	0.7763
30	30.75		0.7025
30	34.08		0.6357
30	37.41		0.5756
31	31.42		0.088

Table 2.3: Multi-point LOD scores observed for chromosomes where a LOD score of greater than zero was observed. Significant LOD scores are highlighted orange and the corresponding position and marker shown in bold text.

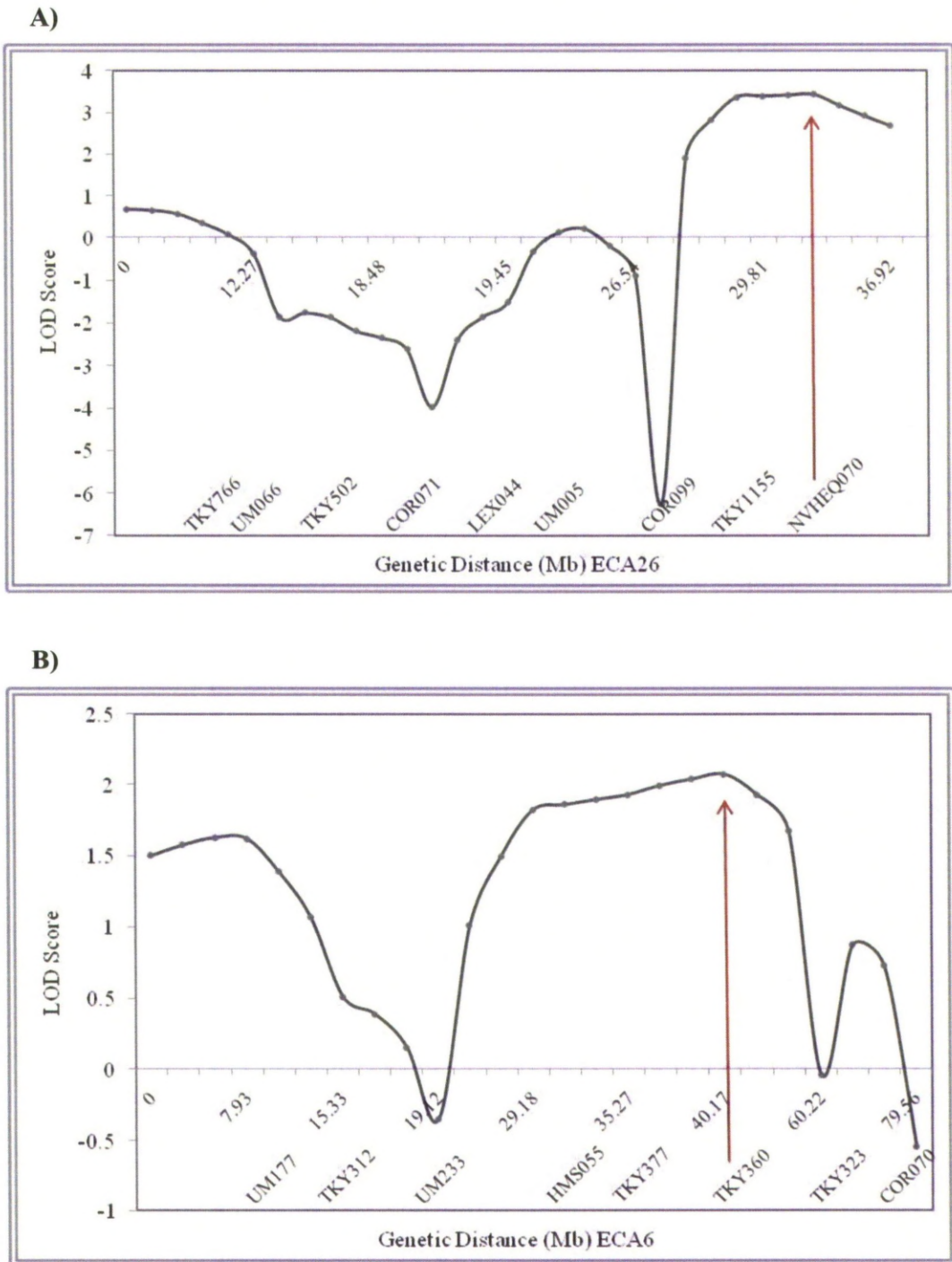


Figure 2.5: Plot of the chromosomes which showed a LOD score of greater than one with multipoint mapping. **A:** LOD scores obtained for nine markers distributed across ECA26; maximum LOD score of 3.42 **B:** LOD scores obtained for eight markers distributed across ECA6; maximum LOD score of 2.07. The red arrows indicate the markers that showed the highest LOD score.

2.4 Discussion

The aim of this study was to perform a genome-wide microsatellite scan in order to identify a chromosomal region which showed a significant linkage to FIS, which could then be carried forward for further analysis. Pedigree analysis strongly suggested that this disease has an autosomal recessive mode of inheritance and therefore only one locus will be responsible for the FIS phenotype. Two approaches to the analysis of the data were taken; the first was a traditional linkage analysis and the second a homozygosity mapping approach where significant differences in homozygosity between the cases and controls were sought.

In the two-point linkage analysis LOD scores greater than one were observed for markers (n=21) on 16 chromosomes and on two chromosomes with multi-point mapping. Statistically significant linkage is acknowledged when the LOD score exceeds 3; based on this, ECA26 was considered the most promising chromosome as the location of the FIS mutation. In addition, homozygosity mapping revealed that marker NVHEQ070 on ECA26 had a significant difference in allele frequencies between the cases and controls. Marker NVHEQ070 was the only marker used in this whole-genome study that showed significant disease linkage, and furthermore this was consistent in both two-point and multi-point linkage analysis and in the homozygosity analysis.

The results of the study do strongly suggest that ECA26 is the most likely location of the mutation as marker NVHEQ070 was the only marker to consistently show an linkage to the disease in all of the analyses. Not only was this marker consistent but it also displayed the highest association in both the two-point and multipoint linkage analysis and in the homozygosity mapping. Based on the linkage studies, the disease linked region spans a ~11.97 Mb region (~28.28 – 40.25 Mb), on a chromosome which is 41.87 Mb long. Therefore the entire chromosome from ~28.28Mb onwards should now be further interrogated by fine mapping.

Due to the relatively small dataset used for this study, the power to detect significant linkage was greatly limited. This is because high heterozygosity, high allele numbers and large sample sets are needed for sufficient power to detect linkage. The Fell Pony, a rare breed which has relatively few founders and has undergone recent genetic bottlenecks, has reduced genetic variation and therefore a higher chance of homozygosity at each marker, so reducing the power of this study. Therefore, none of the loci used in this study that exhibited positive linkage can be disregarded conclusively and further studies are required to definitively confirm the single disease locus. Furthermore, the limited number of samples used would have had a detrimental effect on the power to detect linkage. At the time this whole-genome scan was undertaken, all of the available samples were used. In an attempt to overcome low sample numbers, a homozygosity mapping approach was also taken as it has been successfully used to map disease loci using limited sample numbers (Tryon et al., 2007). The homozygosity mapping conclusively mapped the disease loci to chromosome 26 as no other marker showed a significant difference in allele frequencies between the affected and unaffected group.

When assessing disease linkage, the marker panel should be considered carefully in terms of coverage of the genome. This is because insufficient coverage of the genome can provide the opportunity for error, as the true disease loci could be overlooked because there are insufficient markers to detect linkage with the disease locus. The average spacing of the markers used for the linkage analysis was 9.01Mb, which should provide sufficient coverage of the genome to reliably detect disease linkage. However, the largest marker interval was in excess of 40Mb so some areas of the genome may have been inadequately assayed, and the FIS loci missed.

The aim of this study was to definitively identify a chromosomal region where the FIS mutation was most likely to lie, by performing both linkage and homozygosity based studies. Although this investigation did not conclusively reveal the location of the FIS mutation, it did suggest that chromosome 26 was a highly probable location for the disease mutation. Further studies were planned to either confirm or disprove disease linkage on chromosome 26. Additional sample collection which had been undertaken during this study and the recent release on the EquineSNP50 Beadchip (Illumina, Inc) will enable a genome-wide association study to be undertaken to definitively identify the associated locus. This approach will interrogate thousands of

Single Nucleotide Polymorphism (SNP) across the genome, providing uniform and dense coverage of the genome. This approach has proven especially successful when mapping simple Mendelian diseases in other species (Karlsson et al., 2007) and is the subject of the following chapter (Chapter 3).

Chapter 3:
Confirming the associated chromosome using SNP GWAS

	Page
Summary	55
3.1 Introduction	55
Recent advances in the equine genome	55
Association studies for detecting disease loci	57
Analysis of genetic association data – considering the study design	58
The basic principles of performing GWAS using the Illumina Beadchip	61
Objectives of the FIS disease mapping association study	63
3.2. Materials and Methods	64
Animals and Samples	64
DNA extraction from tissue and blood samples	65
Single Nucleotide Polymorphism Genotyping using the Beadchip	67
3.3. Results	71
Additional sample collection for this investigation	71
SNP genotyping analysis	76
Assessing population stratification	78
Genome-wide association mapping	80
Linkage disequilibrium of the FIS-associated region and haplotype based association test	85
3.4. Discussion	89

Summary

The whole-genome microsatellite scan identified significant FIS linkage on chromosome 26. However, due to the relatively small marker and sample numbers used, disease linkage could not be definitively confirmed as the panel of microsatellite markers did not provide comprehensive cover of the genome. It became imperative to confirm (or deny) this conclusion by some other means. Fortunately, in recent years the approach of whole genome SNP analysis has become available which can be used to compare normal and disease cohorts, at many levels and in many species. Sequencing of the horse genome led to the identification of over one million equine Single Nucleotide Polymorphisms (SNPs), which in turn led to the development of an equine SNP microarray. The microarray lends itself well to genome-wide association studies, providing an excellent alternative for investigating disease genes as it covers the genome far more effectively than microsatellites can achieve. A SNP-association based approach was adopted here to supplement previous findings and definitively confirmed an association between ECA26 and the FIS phenotype. This chapter will introduce genome-wide association studies and present the results of mapping the FIS locus using this approach.

3.1 Introduction

Recent advances in equine genomics

The Horse Genome Project (<http://www.uky.edu/Ag/Horsemap/welcome.html>) was formed in October 1995 and since, has produced several generations of analytical and diagnostic resources for disease mapping in the horse. Until recently, whole-genome disease mapping efforts could only be made using microsatellite markers, a method which has successfully led to the identification of many gene traits (Brunberg et al., 2006, McCue et al., 2008). In July 2005, the equine genome was selected by the National Human Genome Research Institute (NHGRI) as one of 24 mammalian genomes to be fully sequenced to help further understand the human

genome. Mammals were selected to give a diverse overview of the phylogenetic tree. The horse is one of three species from the perissodactyla order and because no other species from this order had been sequenced, the horse was selected. In February 2006, sequencing of the equine genome commenced and by January 2007, the first assembly (EquCab1), which gave a 6.8X coverage of the genome, became publicly available (<http://www.nih.gov/news/pr/feb2007/nhgri-07.htm>); in September 2007, the second assembly was released (EquCab2). The second assembly not only provided increased coverage of the genome, rising from 84% to 95%, but also rectified some of the assembly errors from the first assembly, one of which was that chromosome 26 had been anchored upside down. Both assemblies can be accessed through the following public databases: UCSC Genome Browser (<http://www.genome.ucsc.edu/cgi-bin/hgGateway>) at the University of California at Santa Cruz: The Ensembl Genome Browser (http://www.ensembl.org/Equus_caballus/Info/Index) at the Wellcome Trust Sanger Institute in Cambridge, England: The Broad Institute Web site (<http://www.broadinstitute.org/ftp/pub/assemblies/mammals/horse/>). The genome sequence for the horse comes from a Thoroughbred mare, which was chosen because she was inbred with extensive tracts of homozygosity. Once the Thoroughbred genome sequence was established, several horse and pony breeds were selected to mine for polymorphisms and assess linkage disequilibrium (LD) within and across breeds. These studies led to the identification of approximately 1 million SNPs. By analysing these SNPs in and across breeds, it was established that LD initially drops off quickly to within two-fold of the background level by 100-150Kb, but levels off at 1Mb and remains above background level for ~2Mb (Wade et al., 2009). Importantly, identification of these SNPs facilitated the development of the Equine SNP50 Beadchip (Illumina, San Diego, CA), for genome-wide association studies (GWAS). The Equine SNP50 Beadchip features 54,602 highly informative SNPs uniformly distributed across the entire genome (average spacing of ~43.2 Kb).

Association studies for detecting disease loci

Genetic association studies aim to detect a statistical relationship between one or more genetic polymorphisms and a phenotypic trait. Association studies can take one of two approaches (Fig 3.1). The first is the direct approach, whereby polymorphisms in candidate genes are examined as possible causal variants (Cordell and Clayton, 2005). An issue with direct association studies is the difficulty in identifying plausible candidate variants, particularly when a disease is novel and also when the genome sequence is not comprehensively annotated. The second approach is an indirect study in which a panel of markers are screened for linkage disequilibrium with the causal variant. Coverage of the genome is extremely important in an indirect study as there must be sufficient markers to adequately assay the genome, otherwise the possibility of a causal variant not being identified exists. The Beadchip lends itself well to the indirect approach, enabling a genome-wide scan to be performed at a relatively low cost, with sufficient genome coverage so that the chance of an association being overlooked is greatly reduced.



Figure 3.1: *Approaches for testing for an association using SNPs. A) Direct association method where a candidate SNP (red) is tested for an association with the gene of interest (green box). SNPs are selected based on disease aetiology and gene function. B) An indirect association test utilises LD to assess for an association. The red SNPs are in LD with the causal (blue) SNP. Image reproduced from Hirschhorn and Daly, 2005.*

Analysis of genetic association data – considering the study design

Analysis of the data depends crucially on which of two study designs is adopted: Either the family-based transmission disequilibrium test (TDT) or the population case-control study approach.

Association study results are typically displayed in a Manhattan plot, with the x-axis showing the SNPs on the individual chromosomes and the y-axis indicating the measure of the probability that a variant is associated with that trait.

Case-control studies

Case-control studies compare the frequency of SNP alleles in each group: cases have a robust phenotype for the disease being studied, and controls are known to be unaffected by the disease or have been randomly selected from the population (known as unselected controls) (Lewis, 2002). An increased allele frequency in the cases when compared to the control population may indicate linkage disequilibrium between the typed marker and the causal variant.

Human studies have shown that sufficient sample size is crucial to reach the power needed to detect an association. One approach to maximize power when affected sample numbers are limiting is to use unequal case control ratios, by increasing the number of controls. The power increase will not be as high as when increasing both case and control numbers but may give sufficient increase to detect an association (Spencer et al., 2009). Genome-wide case-control association studies have successfully mapped Mendelian traits in several species using relatively few numbers. In the dog, disease traits have been mapped using 10 cases and 10 controls (Karlsson et al., 2007) and in the horse as few as six cases and 30 controls (Brooks et al., 2010) have yielded success.

Population stratification – A potential problem in case-control association studies

Spurious associations can arise when cases and controls come from populations that are ancestrally different and so have naturally occurring differences in allele frequencies. This is because a case-control association test assumes that any

difference in allele frequency between the two groups is solely due to the phenotype under investigation.

An obvious approach to overcome population stratification is to carefully select cases and controls, matching them on the basis of genetic background. However this is not always possible and so methods have been developed to test and correct for population stratification, to overcome the dangers associated with this. These methods attempt to detect and correct for stratification, minimising the effect of spurious associations whilst maximising the power to detect a true association. A widely used approach is to infer population structure and then incorporate this into the analysis, essentially testing for an association within the identified sub-populations. Many methods exist to assess for and correct population structure but some of the most widely used approaches include multidimensional scaling (MDS) of identity-by-state distances and principle component analysis (PCA) using software such as EIGENSTRAT (Price et al., 2006). However, most of these methods have been developed specifically for human based association studies so a potential risk is that they may overcorrect for stratification in domesticated animal studies, which would result in a loss of power. Furthermore, these methods have limited power when accounting for other types of relatedness, such as family structures (Price et al., 2010) and therefore when performing association studies in domesticated animal species, where potentially cryptic relatedness is likely to exist, other approaches such as family-based studies may offer a solution.

Family-based association studies

Family-based association studies offer an excellent alternative to the case-control association test for association studies where family relationships are known to, or are likely to exist, as is such with inbred domesticated species. Not only are family-based association studies immune to the effects of stratification (Price et al., 2010) but significant findings always imply an association in the presence of linkage (Laird and Lange, 2006). The transmission disequilibrium test (TDT), is the most common family-based association study method and is the one from which most other family-based methods are derived. The TDT test uses genotyping data from trios (sire, dam and affected offspring), capitalising on the Mendelian principle that for any polymorphic marker, each parent contributes one allele to its offspring. The TDT

simply tests for distortion in allele transmission from the heterozygous carrier parent to the homozygous affected offspring (Lewis, 2002). The TDT test is the simplest of the family-based methods, using very little computational power, but is limited in its use as it can only be used on nuclear families where all three animals have genotypes. Additionally, it does not allow other family members and distant relatives to be used in the analysis. To overcome this, many extensions of the TDT have been developed to enable family-based association studies to be carried out on extended families with missing individuals. Further to this, some of these methods also enable genotyping data of unrelated animals to be incorporated into the analysis by inferring family relationships via cluster analysis.

Family-based studies are not only an attractive alternative to case-control studies because they are immune from stratification but also because they are an extremely useful approach to use when following up on linkage based studies where samples from family groups have already been collected. However a potential issue is that sufficient samples with accurate pedigree information need to be collected to provide sufficient power to detect meaningful associations, something which can be difficult when pedigree data is sensitive or inaccurate.

Haplotype analysis for added power to detect an association

Haplotypes refer to a group of alleles which are located within close proximity to one another on a chromosome and so tend to be inherited together. Haplotypes are extremely important in association studies as they enable us to understand the pattern of LD across the genome (Liu et al., 2008). The power to detect an association using a single-marker based method is limited to the LD between the marker and the disease locus. Based on this, haplotype based association studies are generally considered to be more powerful than the single-marker based approach as it incorporates LD information contained in multiple markers (Cordell and Clayton, 2005, Jin et al., 2010). The simplest method to infer haplotypes is to use a predefined sliding window (e.g. a set number of SNPs in the haplotype) or alternatively, more complicated approaches can be used such as computational methods, for example cluster analysis based haplotypes (Tzeng et al., 2006).

Multiple testing in genome-wide association studies

A fundamental issue with the analysis of case-control and family-based studies, is the interpretation of the results. This is because due to the large number of statistical tests performed, a large number of associations will undoubtedly be spurious (type 1 error). The most widely accepted approach to overcome the problem of multiple-testing is to adjust the genome-wide significance level for each test (Moskvina and Schmidt, 2008). The simplest way to do this is by using the Bonferroni correction, adjusting the level of significance to a discretionary level, although it widely accepted that a significance level of 5% is suitable ($0.05/\text{total number of markers used in the analysis}$). However, these simple methods do not account for LD amongst the SNPs which can result in an overly conservative P -value (Pahl and Schafer, 2010).

Permutation based (Churchill and Doerge, 1994) methods generate significance levels empirically, so fully account for LD amongst the SNPs and are therefore the most widely used and accepted method for accounting for multiple-testing correction in GWAS.

The basic principles of performing GWAS using the Illumina Beadchip

Isolating and preparing DNA for genetic investigation

Genomic DNA may be isolated from blood and/or tissue samples for SNP genotyping. For the principles of isolating gDNA, see Chapter 2 (Isolating DNA for genetic Investigation). A minimum of 750ng of gDNA is required for the genome-wide SNP assay (15 μ l at 50ng/ μ l).

An overview of the Illumina EquineSNP50 Beadchip

The EquineSNP50 Beadchip uses the Infinium HD assay chemistry to interrogate 54,602 SNPs. This chemistry uses a two-step approach, in which carefully selected 50-mer probes selectively hybridise to the loci of interest, stopping one base before the base for interrogation. The second stage involves single-base extension,

incorporating a labelled nucleotide which can then to be deduced by an image processor (Fig 3.2). This powerful and accurate chemistry yields call rates >99%, has reproducibility >99.9% and has <0.1% errors assessed from Mendelian inconsistencies

(http://www.illumina.com/products/equine_snp50_whole_genome_genotyping_kits.ilmn).

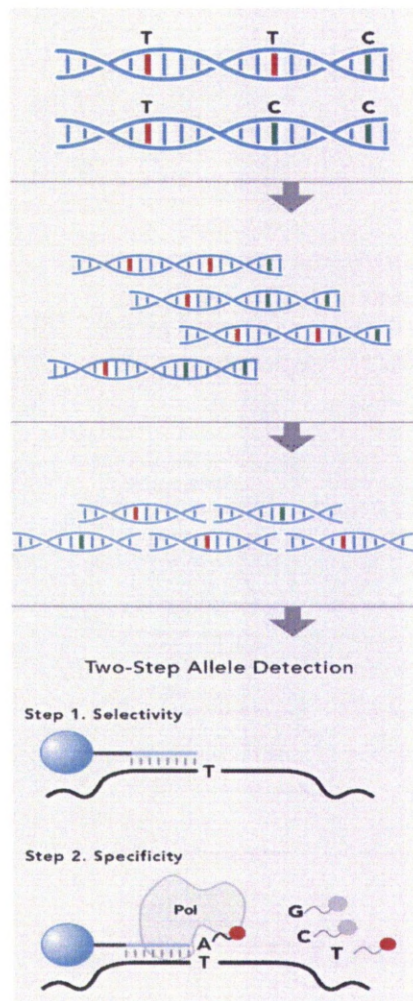


Figure 3.2: The two-step approach used for the Infinium HD assay. Initially PCR-free whole-genome amplification of the gDNA is performed. The amplified DNA is then fragmented and added to the Beadchip where the fragmented DNA hybridises to the corresponding 50-mer probe that is attached to the array. Finally, enzymatic single-base extension occurs with a labelled nucleotide which can be deduced by an imaging processor. Image reproduced from Illumina, Inc (http://www.illumina.com/technology/infinium_hd_assay.ilmn).

A potential problem which has been identified in previous GWAS experiments is that differences in laboratory treatment of samples, the use of multiple Beadchip batches and processing samples at different times, can result in differential bias (Clayton et al., 2005). Therefore, it is essential that all DNA samples are handled and treated in the same way and that all samples must be processed together using the same Beadchip batch.

Objectives of the FIS disease mapping association study

SNP based GWAS has proven to be very successful when mapping recessive traits in domesticated species (Charlier et al., 2008, Awano et al., 2009). Therefore, with the addition of more samples, an indirect association study approach using the EquineSNP50 Beadchip was used to map the FIS disease loci and either confirm or disprove the previously identified association on ECA26. The genotyping data was analysed using both the case-control association test and a family-based association test using PLINK (Purcell et al., 2007). With both analyses, the data was corrected for multiple testing using the permutation method, to establish if any of the observed associations reached the threshold for statistical significance. In addition to this, by performing the two different analyses, a comparison of the two approaches could be made in terms of their use in domesticated animal studies.

This chapter will describe -

1. Experiments designed to definitively map an associated locus with the FIS phenotype using the EquineSNP50 Beadchip.
2. The SNP data will be further mined and analysed to estimate the boundaries of the associated chromosomal region for subsequent fine-mapping studies.

3.2. Materials and Methods

Animals and Samples

Fifty-three blood or tissue samples (appendix 1) were selected for the SNP genotyping. Of these, forty-six were collected as part of Dr. Gareth Thomas's PhD studies (Thomas, 2003) consisting of seventeen affected Fell pony foals and twenty-nine controls (adult Fell ponies). The remaining seven samples were collected by Mr. P. May, a veterinary surgeon in private practice (Newbiggin, Cumbria), during this PhD study and consisted of five affected foals (tissue samples) and two adult Fell Ponies (blood samples). All samples were collected under the Veterinary Surgeons Act 1966. Procedures in living animals were limited to the collection of jugular venipuncture, inflicting minimal, if any pain. Disease status for affected foals was confirmed post-mortem based on gross findings, haematological and histological findings (Scholes et al., 1998).

Histological examination

Tissue samples were collected post-mortem from suspected FIS-affected foals and immersed immediately in 10% formalin. Tissue samples were obtained from the spleen, thymus, lymph nodes and bone marrow. Samples were transported to the Animal Health Trust for histological and immunohistochemical analysis; sections were prepared by Ms. N. Flindall and examination conducted by Professor, T. Blunden, at the Animal Health Trust, Newmarket. Histological staining included haematoxylin and eosin (H&E) to assess the general histological architecture of tissues, Perl's Prussian blue for ferric iron, and immunohistochemical staining for CD3+ (marker for T-lymphocytes) and CD79A (marker for B-lymphocytes).

Sample collection by Dr. Gareth Thomas

Jugular blood samples were collected by Dr. G Thomas or by a veterinary surgeon in private practice and posted to the University of Liverpool. Samples were collected into Vacutainers® containing ethylenediamine tetra-acetic acid (EDTA)

anticoagulant (Beckton Dickinson, UK). Samples were archived and frozen on receipt at -20°C.

Tissue samples (kidney, spleen, muscle and lymph node) were collected during post-mortem examination from some of the affected foals at the Veterinary Investigation Centre (Merrythought), Penrith or at the Department of Veterinary Pathology, University of Liverpool. Fresh tissue samples were frozen and archived on receipt at -20°C.

Additional sample collection performed during this study

Jugular blood samples were collected by Mr. P. May, a veterinary surgeon in private practice (Newbiggin, Cumbria). Samples were collected into 10mL ethylenediamine tetra-acetic acid (EDTA) Vacutainers® (Beckton Dickinson, UK). On receipt samples were frozen at -20°C and archived.

Tissue samples (kidney) were collected during post-mortem examination from all affected foals by Mr. P. May, a veterinary surgeon in private practice (Newbiggin, Cumbria). On receipt samples were frozen at -20°C and archived.

DNA extraction from tissue and blood samples

DNA extractions by G. Hill

Twenty-six of the extractions were performed by Ms. G. Hill using Nucleon® gDNA extraction kit (GE Healthcare, UK). DNA was extracted according to the manufacturer's instructions and re-suspended in ddH₂O. All samples were further processed using the MultiScreen PCR₉₆ 96-well filter plates (Millipore®, USA), so the samples could be eluted in TE buffer (Sigma-Aldrich, UK).

Additional DNA extractions

Genomic DNA was isolated from blood using a revised version of the Nucleon® blood extraction kit and isolated from tissue samples also using a revised version of the Nucleon® blood extraction kit (GE Healthcare, UK). Protocols for the two extraction methods can be found in Appendix 2 and 4. DNA was extracted from 1ml

of blood and 150mg of kidney tissue. DNA was re-suspended in TE buffer (Sigma-Aldrich, UK).

DNA Quality control procedures

All DNA samples were assessed for quality using a Nanodrop® spectrophotometer and for quantity with Picogreen® reagent (Invitrogen, USA). Samples where a low concentration or contamination was observed were purified and concentrated using MultiScreen PCR₉₆ 96-well filter plates. For SNP genotyping, the minimum gDNA requirement is 750ng. Samples were normalised to a target concentration of 60ng/ul for genotyping.

Preliminary analysis of sample concentration using the Nanodrop®.

Initially, sample quality was assessed using a Nanodrop® by measuring the 260/280 and 260/230 ratios. Samples were processed in accordance with the manufacturer's instructions (<http://nanodrop.com/Library/CPMB-1st.pdf>).

Samples where a 260/280 ratio of <1.8 or a 260/230 ratio <2.0 was observed were processed using filter plates, to remove contaminants. After processing, samples were re-assessed using a Nanodrop® and samples that still showed evidence of contamination were discarded and the DNA extraction from the original sample repeated.

Quantification of double stranded DNA using Picogreen®

Picogreen® analysis was subsequently used for quantification of double stranded DNA (Singer et al., 1997) in the samples to be SNP genotyped. Although measuring absorbance with a Nanodrop® provides an estimate of the concentration of nucleic acids, the major disadvantage of this method is that the concentration measurement also includes single-stranded DNA and any contaminants that may be present. Therefore, to give an accurate measurement of the concentration of double-stranded DNA, Picogreen® was used as it fluoresces only on binding to double-stranded DNA, enabling an accurate measurement to be taken.

Samples were processed using the Quanti-iT™ Picogreen® dsDNA reagent kit, in accordance with the manufacturer's instructions (<http://probes.invitrogen.com/media/pis/mp07581.pdf>). Sample fluorescence was

measured at a single time-point using a Techne Quantica® Real-Time Cycler. Three replicates of each sample were prepared and the DNA concentration (ng/ul) calculated as the average of the three replicates.

Purification and quantification of DNA samples

MultiScreen PCR₉₆ 96-well filter plates were used for samples where contamination or low concentration had been observed. Samples were diluted to 200ul using ddH₂O and transferred to the wells of the Multiscreen plate. The plate was placed on a vacuum manifold at 24inch mercury for 20 minutes. Appropriate volumes of TE buffer were then added to the individual wells, incubated at 37°C for 10 min on a tilting platform, and then transferred to individual storage tubes.

Single Nucleotide Polymorphism Genotyping using the Beadchip

Genotyping using the EquineSNP50 chip was performed by The Cambridge Genomic Services Laboratory, Cambridge University (Pathology Department).

Sample processing – An overview of processing the Beadchip

The equine Beadchip is a commercially available product which interrogates 54,602 SNPs in one single assay (Fig. 3.3). In the standardised procedure, the gDNA samples are amplified overnight (~20 hours) by isothermal amplification. The amplified product was then fragmented by enzymatic action, precipitated in alcohol, re-suspended and then applied to the Beadchip. During an overnight incubation, hybridisation of the sample to the Beadchip occurs as the DNA anneals to locus-specific probes which are covalently linked to one of the 53,602 beads. After hybridization, fluorescently labelled enzymatic single base extension occurred which was then in turn detected by the Illumina Beadchip reader, producing intensity files for each bead. The intensity data was then imported into GenomeStudio™ for automated genotype calling.

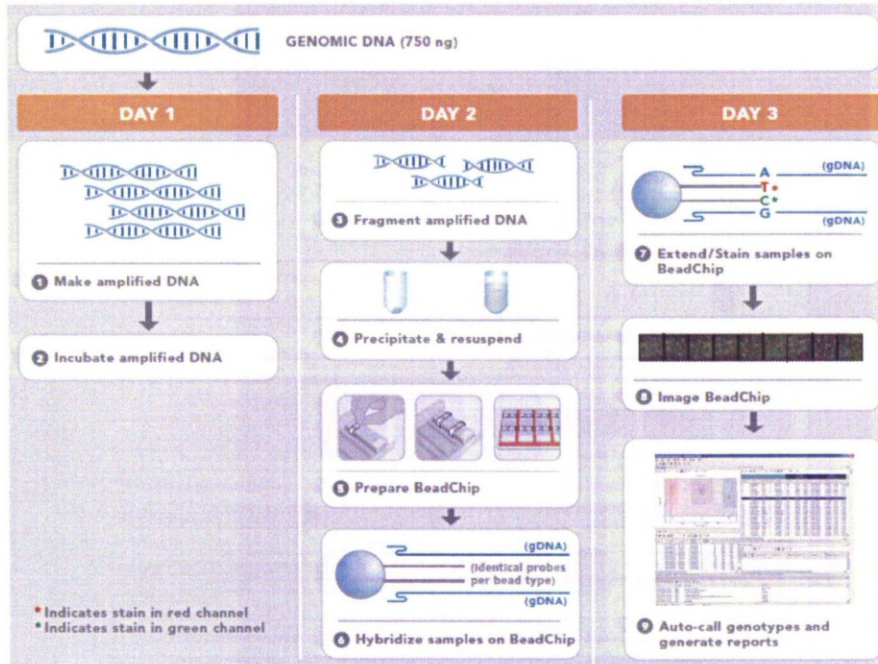


Figure 3.3: An overview of the Infinium II whole-genome genotyping assay, showing the 3 day, 9 step protocol. Image reproduced from Illumina, Inc: (http://www.illumina.com/Documents/products/workflows/workflow_infinium_ii.pdf)

Analysing Infinium genotyping data

Data files were uploaded into the GenomeStudio™ Genotyping Module (Illumina inc, USA) and clustered using a cluster file which had been generated from 1400 Thoroughbred horse samples which were processed alongside the Fell Pony samples, creating a project for visualisation of SNP genotypes. Uploaded files included the SNP manifest file (.bpm) which provided detail of the SNPs on the array, the intensity files (.idat) which were generated when the array was scanned, a cluster file (.egt) and a sample sheet (.csv) which contained sample specific phenotype information (sex and affection status).

Initially, automatic gender estimation was performed in GenomeStudio™ and compared to samples where gender was known. This was used as a check for processing errors. Samples that performed poorly were identified with a call rate of <95% and excluded from further analysis. SNP clusters were then edited manually in accordance with the ‘Infinium® Genotyping Data Analysis Manual’

(http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf) using steps 1-9 in the 'Evaluating and Editing SNP cluster position' guide, excluding step six which could not be applied to this dataset: 1) Evaluating SNPs with overlapping clusters. 2) Evaluating SNPs with a low call frequency. 3) Evaluating SNPs with low intensity data. 4) Evaluating cluster separation. 5) Identifying Mendelian inheritance errors within nuclear families. 6) Evaluating reproducibility errors. 7) Assessing SNPs with excess heterozygotes. 8) Evaluating loci where homozygotes have been incorrectly identified. 9) Assess heterozygous males for X-linked markers.

After completing the manual cluster editing, sample statistics were calculated in GenomeStudio™ and the final project saved ready for downstream analysis. The final report (.txt file) was imported into PROGENY (Progeny Software LLC, South Bend, IN, www.progenygenetics.com). Pedigree data was also imported into PROGENY for samples with known pedigree information, forming one extended pedigree. Finally, PROGENY was used to create the files (.map and .ped) for statistical analysis in PLINK.

Statistical analysis of the genotyping data - PLINK

Initially the dataset was filtered to remove SNPs with a minor allele frequency (MAF) $\leq 2\%$ and SNPs with $\geq 10\%$ missing genotypes using PLINK. All subsequent analysis was performed on this filtered dataset.

Population stratification was assessed using PLINK. Initially, pairwise identity-by-state (IBS) distances were calculated (--ibs-test) to identify how statistically different the case and control populations were. Multidimensional scaling (MDS) analysis on the $N \times N$ matrix of genome-wide IBS pairwise distances was performed (--mds-plot --cluster) and plotted using Excel to provide a visual representation of clustering as evidence of stratification. A quantile-quantile (QQ) plot (--qq-plot) was used to compare the association statistics under the null hypothesis of no association. The genomic inflation factor was calculated to assess inflation of observed statistics due to relatedness and potential population structure. The plot was also examined visually for a deviation from the null distribution. Deviation from the null distribution by many of the SNPs is suggestive of population stratification whereas a

plot which follows the null distribution until near the end, with a few SNPs deviating, provides evidence of a true association.

A case-control basic association test (-- assoc) was performed within PLINK, using a Bonferroni correction. Additionally, genome-wide significance was ascertained through label-swapping max(T) permutation testing (-- mperm) (T = 10,000). A family-based association test was performed using the family-based association test for disease traits (DFAM) model in PLINK (-- dfam), using the nonfounders option (--nonfounders) to allow for sibling-only analysis (i.e. no parents). DFAM is a sib-TDT based model which allows unrelated individuals to be included (via a clustered-analysis using the Cochran-Mantel-Haenszel test). The clustering analysis infers relationships based on the calculated IBS distances. Genome-wide significance was ascertained through label-swapping max(T) permutation testing (-- mperm) (T = 10,000)

Statistical analysis of the genotyping data - HAPLOVIEW

Statistically associated loci were visualised by zooming in on associated regions and LD plots generated using HAPLOVIEW version 4.2 (Barrett et al., 2005). SNPs from the associated region, were formed into haplotypes using the 'solid spine of LD' option, which searches for a "spine" of strong LD running from one marker to another along the legs of the triangle in the LD chart. Multi-marker haplotype based association tests, corrected for multiple testing with 10,000 permutations were performed in HAPLOVIEW using these haplotype blocks.

3.3. Results

Additional sample collection

Seven samples were collected by Mr. P. May, a veterinary surgeon in private practice (Newbiggin, Cumbria). Two were healthy adult ponies and the remaining five were from Fell Pony foals which were euthanized as they were suspected of being affected by FIS. FIS was confirmed in all five foals post-mortem, based on haematological and histopathological examination.

Clinical examination and haematology

Details of age, sex, packed cell volume and clinical signs observed at the time of euthanasia are summarised in table 3.1. Many of the clinical signs were reported by the owner as persistent from one to two weeks of age. Additionally, many of the foals were reported as unusually calm and easier to handle than other similar aged foals. Several of the foals were also reported as showing inadequate mare bonding, often being left behind by the mare in the paddock.

Foals 05_08 and 09_08 were treated with supportive therapy and antibiotics for in excess of two weeks but showed no significant improvement. All foals were euthanized on the basis of lethargy, anaemia and persistent infections.

Sample ID	Age (weeks)	Sex	Packed Cell Volume	Clinical signs
01_08	10	Filly	9.8%	Respiratory noise, scouring, nasal discharge, good weight, dull but responsive, pale mucous membranes
02_08	6	Filly	4%	Scouring (pale green profuse diarrhoea), abdominal effort with no respiratory noise, pale mucous membranes, weight loss.
03_08	3	Filly	16%	Severe scouring (pale green/yellow watery diarrhoea), moderate nasal discharge, poor weight, dull and depressed, starey coat and mild dehydration.
05_08	6	Filly	22%	History of repeated scouring although no diarrhoea at time of euthanasia, abdominal effort with breathing, dull, nasal discharge and pale mucous membranes.
09_08	8	Filly	Not assessed	No respiratory noise, abdominal effort, very pale mucous membranes, dull but responsive and diarrhoea.

Table 3.1: *Clinical signs and history from five FIS-affected foals, including haematology results and the age at which the foal was euthanized.*

Histology and Immunohistochemistry:

Details of the histological (Fig: 3.4) and immunohistochemical staining can be found in table 3.2 and 3.3. The main observation at post-mortem was that all foals had a distinct lack of thymic tissue; the area in which the thymus would usually be identified, was filled with normal adipose tissue containing thymic ribbons. All histological (Fig: 3.4) and immunohistochemical was performed by Professor T. Blunden, Animal Health Trust, Newmarket.

Sample ID	Histology results
01_08	<p>Lymph node – Lymphodepletion, diminutive follicles.</p> <p>Thymus – Small lobules, no clear corticomedullary demarcation.</p> <p>Spleen – Diminutive splenic follicles.</p> <p>Bone marrow – Erythroid hypoplasia.</p> <p>Liver – Multifocal hepatitis.</p>
02_08	<p>Lymph node – Lymphoid follicles apparent in all sections.</p> <p>Thymus – Small lobules, no clear corticomedullary demarcation.</p> <p>Spleen – Variably sized lymphoid follicles.</p> <p>Bone marrow – Evidence of erythroid hypoplasia.</p> <p>Liver – Tiny multifocal necrotic foci, acute multifocal hepatitis</p>
03_08	<p>Lymph node – Lymphoid hypoplasia, no distinct follicles.</p> <p>Thymus – Possible hyperplasia.</p> <p>Spleen – Diminutive follicles.</p> <p>Bone marrow – Erythroid hypoplasia.</p> <p>Liver – No changes detected.</p>
05_08	<p>Lymph node – Small lymph nodes, lacking distinct follicular architecture.</p> <p>Thymus – Small lobules with no clear corticomedullary demarcation.</p> <p>Spleen – Diminutive splenic follicles.</p> <p>Bone marrow – Erythroid hypoplasia.</p> <p>Liver - No changes detected.</p>
09_08	<p>Lymph node – Lack of distinct follicular architecture.</p> <p>Thymus – Small lobules without clear corticomedullary demarcation</p> <p>Spleen – Diminutive splenic follicles, hypercellular red pulp.</p> <p>Bone marrow – Hypercellular with predominance of granulocyte series, low numbers of late normoblasts.</p> <p>Liver – Evidence of bile duct hyperplasia or fibrosis.</p>

Table 3.2: *Histology results from five FIS-affected foals.*

Sample ID	Immunohistological Results
01_08	Marked reduction in B-lymphocytes, particularly in the lymph node and spleen.
02_08	Reduced B-lymphocyte population, particularly notable in the lymph node, spleen and thymus.
03_08	Reduced B-lymphocyte population, particularly notable in the lymph node, spleen and thymus.
05_08	Marked reduction of B-lymphocytes in all tissues.
09_08	Very diminutive B-cell follicles within the lymph node and spleen, as well as relatively low B-cell population in the thymus.

Table 3.3: *Immunohistochemical analysis of FIS-affected foals.*

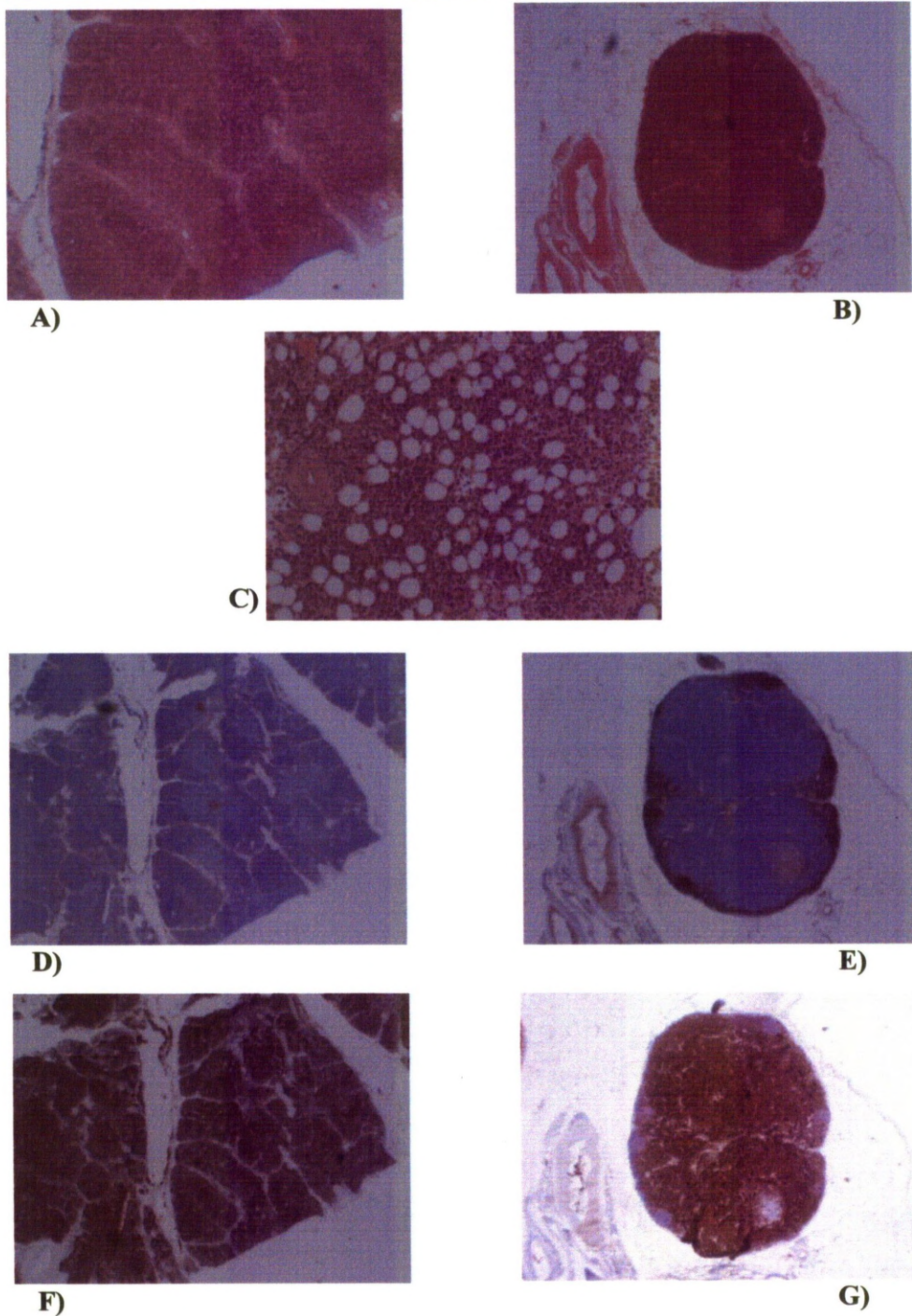


Figure 3.4: *Histological and immunohistological examination of FIS-affected foals. A) Thymus showing small lobules with a loss of distinction between the cortex and the medulla. B) Lymph node showing loss of follicular structure. C) Hypercellular bone marrow. D) Thymus section stained using CD79A, showing a marker reduction in B-lymphocytes. E) Lymph node stained with CD79A, showing diminutive residual follicles and B-lymphocyte depletion. F) Thymus section stained using CD3+, showing T-lymphocytes to be well populated. G) Lymph node stained with CD3+, revealing abundance of T-lymphocytes. All sections shown at 10x magnification.*

SNP genotyping analysis

Of the 53 animals genotyped, consisting of 22 affected animals and 31 controls, four affected animals were excluded from further analysis in GenomeStudio™, as they either had a SNP call rate of <95% (n=2) or phenotype discrepancies had come to light based on additional information (n=2).

Manual editing of genotypes in GenomeStudio™, in accordance with the ‘Infinium® Genotyping Data Analysis Manual’, led to the exclusion of 9,957 SNPs from further analysis. A total of 49 individuals were exported from GenomeStudio™ for further analysis, consisting of 18 affected animals and 31 controls. Pedigree information was imported into PROGENY (Fig 3.5) and an extended pedigree incorporating 34 of the animals was constructed, consisting of 16 affected animals and 18 controls. A total of 15 animals were not incorporated into the pedigree due to missing pedigree information (three affected animals and 12 controls).

Genotypes were further filtered using PLINK: A total of 2,109 SNPs were excluded due to a $MAF \leq 2\%$. After additional SNP pruning, the final number of SNPs used for conducting the association analysis was 42,536 with a mean call rate of 99.6% for the 49 individuals.

Therefore, based on the number of SNPs to be included in the analysis, the level of significance for the basic association test, as determined by a Bonferroni correction ($0.05/42,536$) was 1.17×10^{-6} which is equivalent to a $-\log_{10}$ p-value of 5.92.

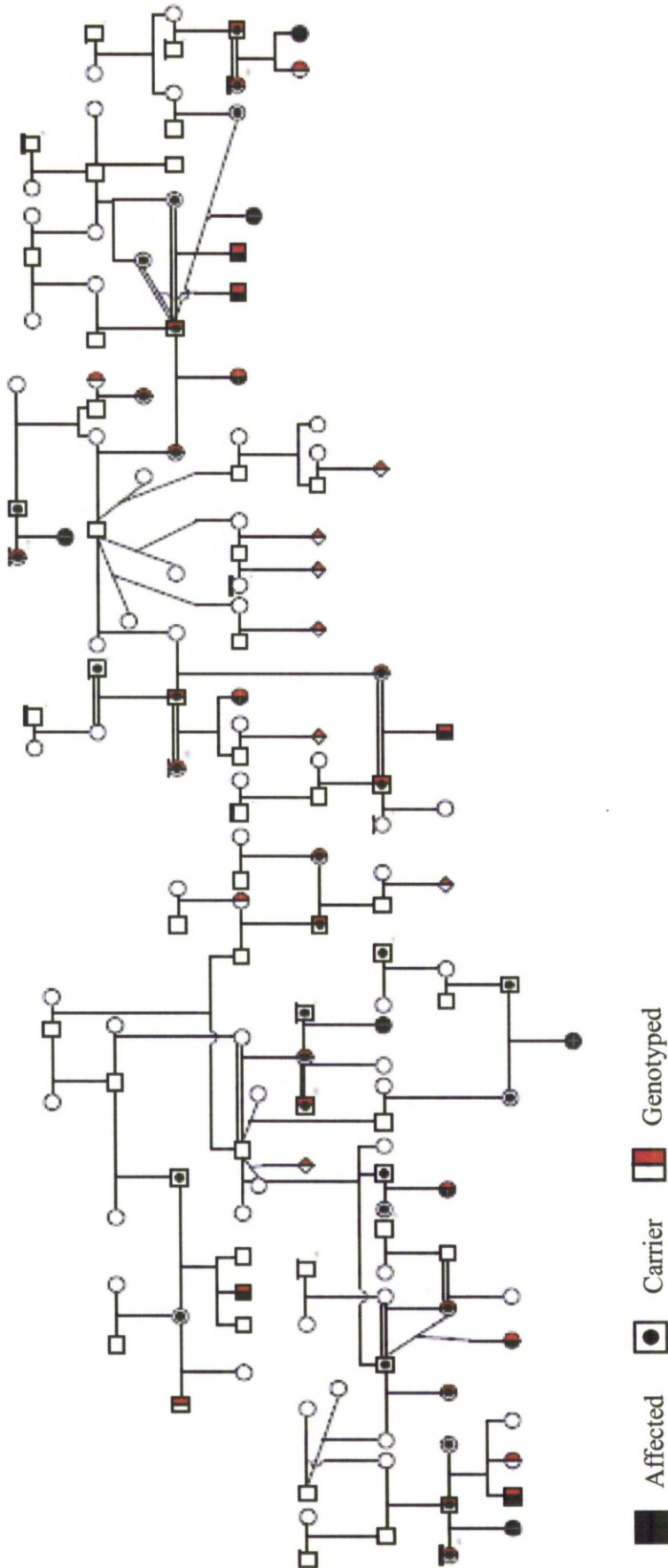


Figure 3.5: Extended pedigree structure constructed in Progeny, showing the relationships between the 34 animals used in the association studies for which pedigree information was available. Samples with a thick black line above them are duplicated on the figure to break inbreeding loops. Samples shown in red were genotyped in the association study.

Assessing population stratification

Prior to association testing, the sample set was assessed for population stratification, which may lead to spurious disease associations. Relatedness of the samples was assessed by calculating identity-by-state (IBS) scores using PLINK. IBS distances between the two groups (cases and controls) showed that they were not significantly different (P -value = 0.55), providing evidence of a closely related population with no apparent sub-structure. Multidimensional scaling plots of the IBS distances revealed that there was no clear separation between the two groups (figure 3.5), providing further evidence of a close relationship between the cases and controls. Furthermore, a Quantile-Quantile plot revealed a genomic inflation factor of 1.04, and visual inspection showed that there was very little deviation from the null distribution (Fig 3.6), indicating that any population stratification is minimal. The sharp deviation above the null distribution line is likely to be due to a strong association of the disease with those SNPs.

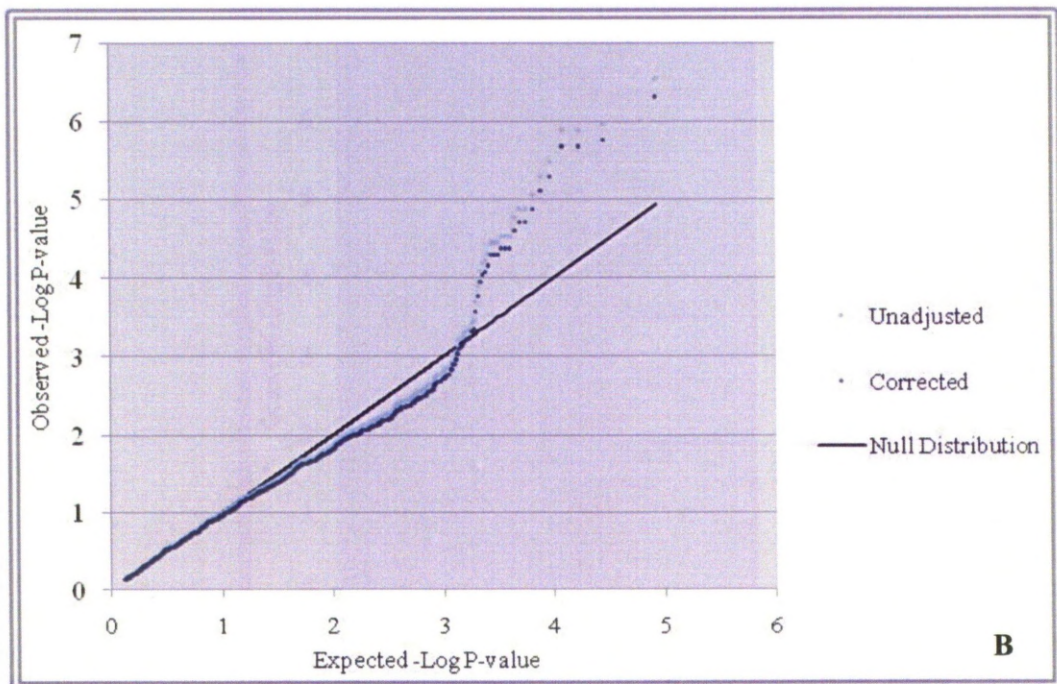
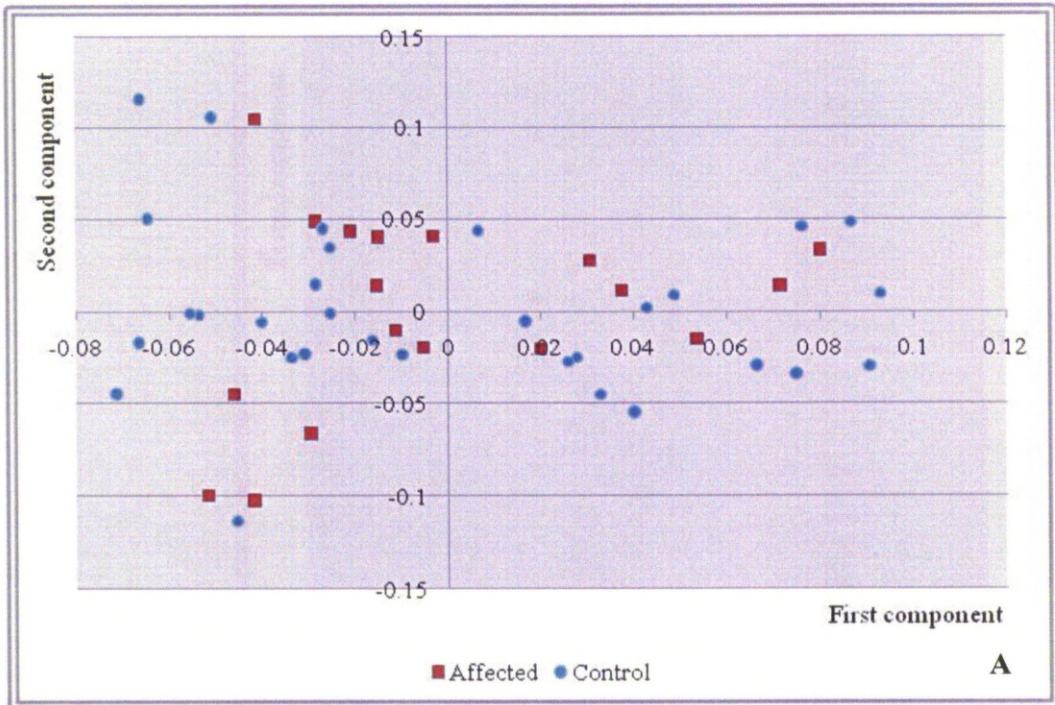


Figure 3.6: Population structure analysis of the control and affected groups: **A)** A multidimensional scaling plot of the IBS distances showing no clear demarcation between the control and affected group. **B)** A QQ plot of observed P-values against expected P-values, shown before and after adjustment for population stratification

Genome-wide association mapping

Case-control association tests were performed in PLINK using 18 affected animals and 31 controls, revealing a single region on ECA26 with genome-wide statistical significance (Fig 3.7). A total of 44 SNPs encompassed the associated region, which spanned 2.92 Mb (ECA26: 29,695,261 to 32,187,972) with two SNPs being statistically significant at $P < 0.05$. Of the 44 SNPs in this region, 13 SNPs were in a contiguous block and homozygous in all of the affected animals (29,803,727 – 30,802,367). The highest FIS association in this region was observed with BIEC2-692674 at 29,804,057 Mb, which displayed a $P\text{-value}_{(\text{raw})}$ of 2.88×10^{-7} . When corrected for multiple-testing with 10,000 permutations, the same 2.92 Mb region on ECA26 was identified as showing an association with the disease (Fig 3.9.A). Five SNPs were identified as genome-wide significant in the corrected association test. The most highly associated SNP was BIEC2-692674 (ECA26) with a $P\text{-value}_{(\text{mperm})}$ of 4.0×10^{-3} .

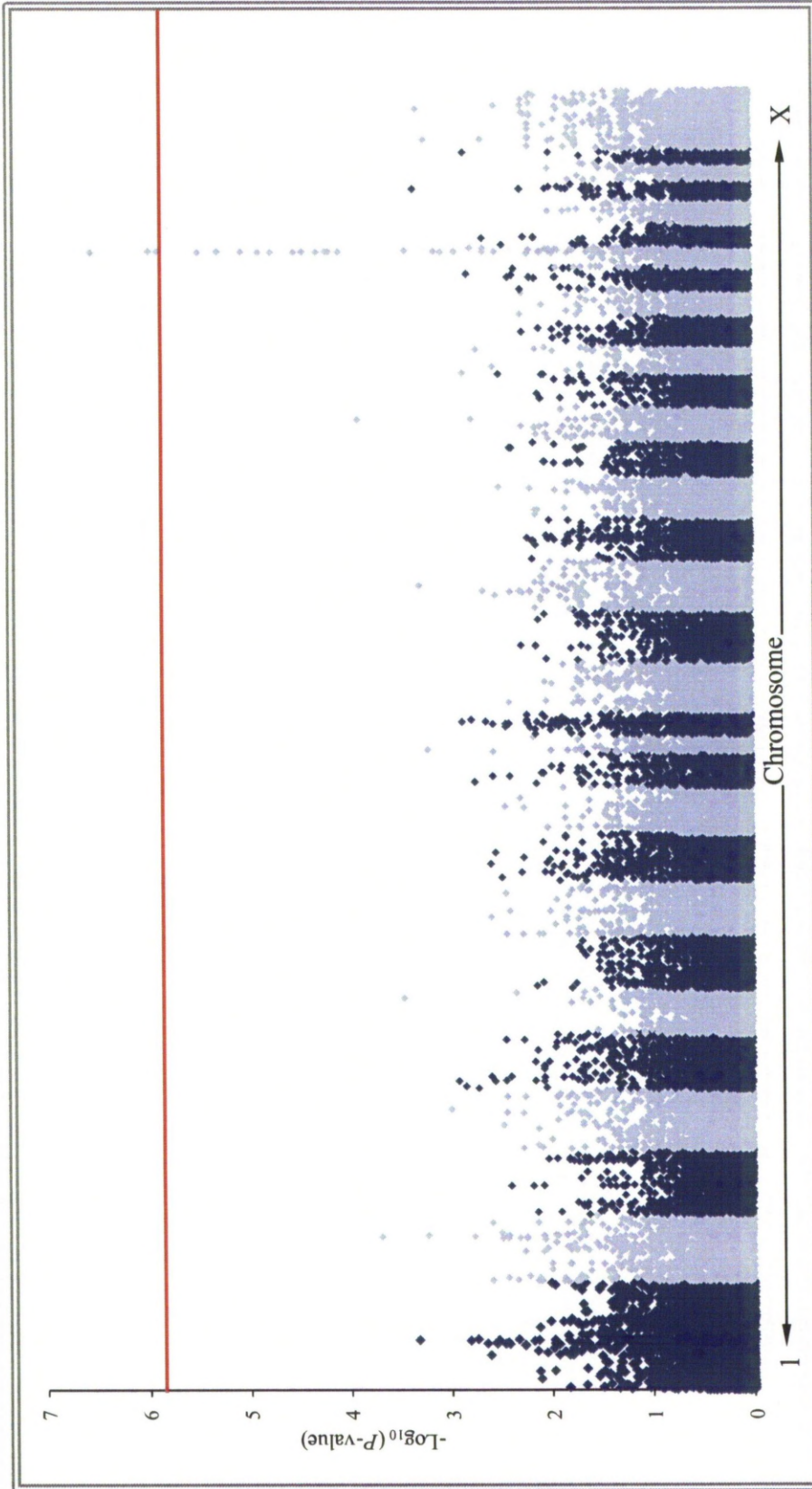


Figure 3.7: Manhattan plot displaying the results ($-\log_{10}$ of p -value) for case-control SNP association with Foal Immunodeficiency Syndrome. Individual chromosomes are represented by alternating colours in numerical order (1-31 and X). The horizontal red line indicates the threshold for genome-wide statistical significance ascertained by Bonferroni correction ($P=0.05$).

A family-based association test was performed using DFAM in PLINK, using 34 samples from an extended pedigree (16 affected animals and 18 healthy relatives) and 15 animals with unknown pedigree data (three affected animals and 12 controls). This identified a single region on ECA26 which was the same 2.92 Mb region as identified by the case-control association approach. The highest association was observed with two adjacent SNPs (BIEC2-693058 at 31,589,364 bp and BIEC2-693062 at 31,593,160 bp) which displayed P -values_(raw) of 4.1×10^{-6} , but neither showed a significant association (Fig 3.8). Both of these SNPs were homozygous in only 14 of the 16 affected animals. When corrected for multiple-testing with 10,000 permutations, the same 2.92 Mb region was associated with FIS, but again the highest association was observed with BIEC2-693058 and BIEC2-693062 with both SNPs displaying P -values_(mperm) of 1.3×10^{-2} (Fig 3.9.B).

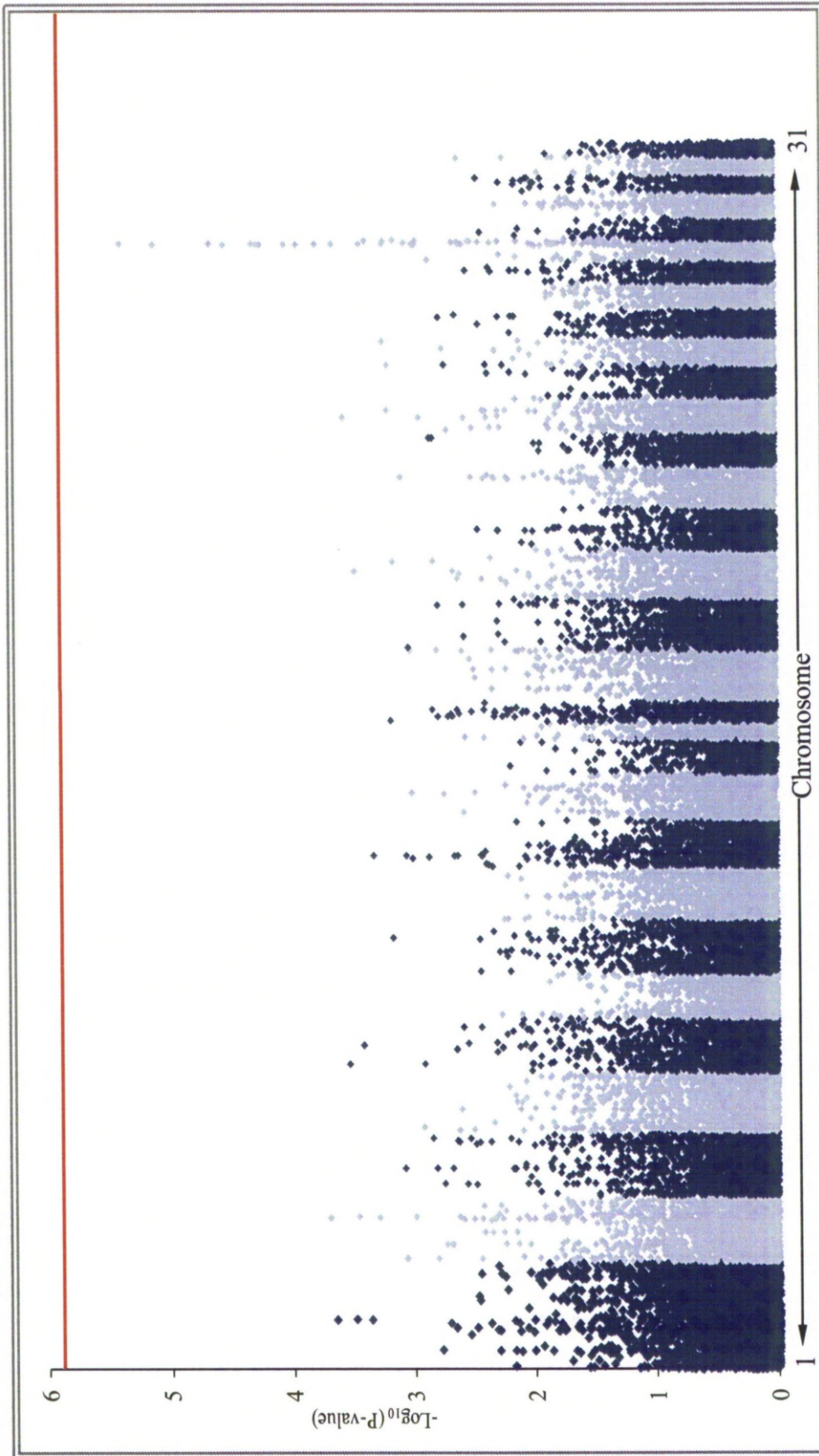


Figure 3.8: Manhattan plot displaying the results ($-\log_{10}$ of p -value) for a family-based SNP association test with Foal Immunodeficiency Syndrome. Individual chromosomes are represented by alternating colours in numerical order (1-31). The horizontal red line indicates the threshold for formal statistical significance at $p=0.05$.

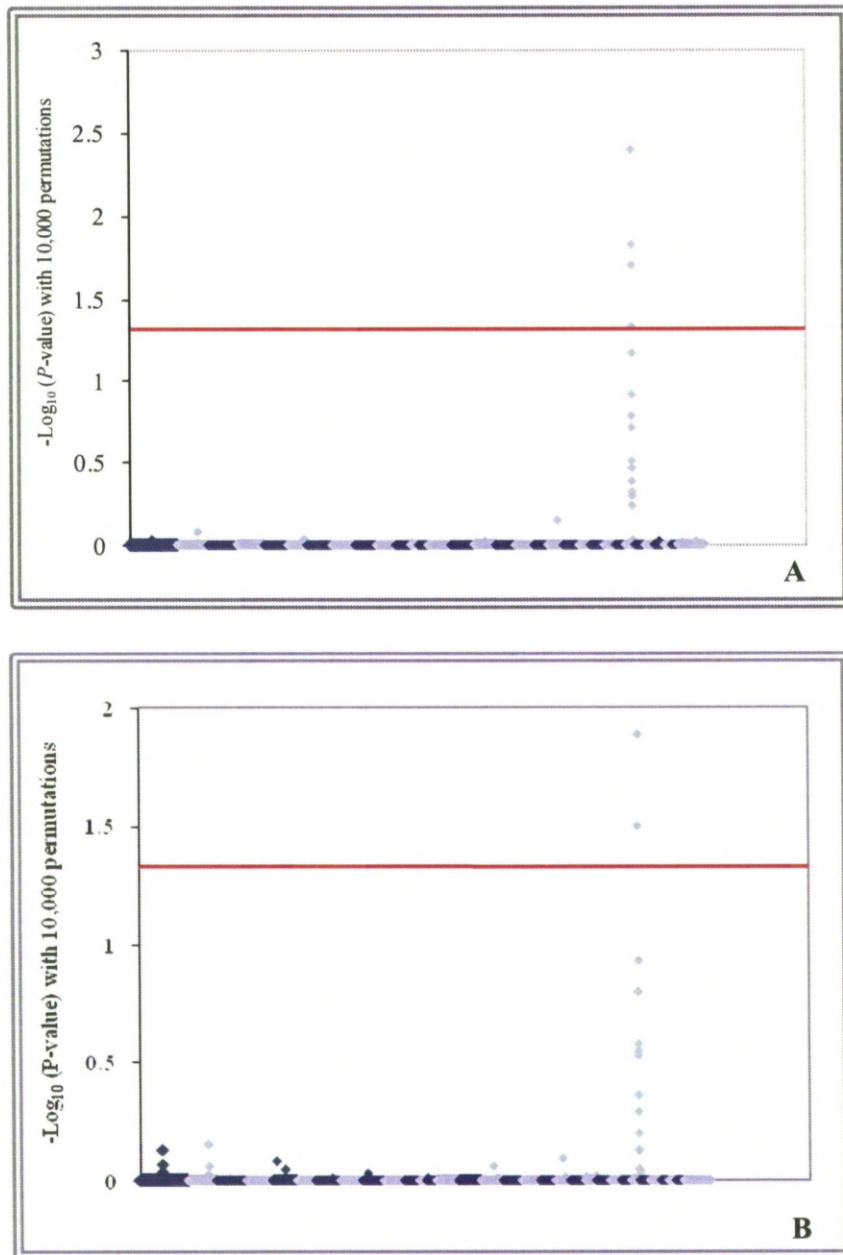


Figure 3.9: Genome-wide mapping of the FIS locus. Significance of the association was calculated empirically with a label swapping approach using 10,000 permutations. Both the case-control and family-based approach identified a single locus with genome-wide significance. Individual chromosomes are represented by alternating colours in numerical order (1-31 and X). The horizontal red line indicates the threshold for formal statistical significance at $P=0.05$. **A)** Case-control association test showed genome-wide significance with 4 SNPs. **B)** Family-based association test showed genome-wide significance with 2 SNPs.

Linkage disequilibrium of the FIS-associated region and haplotype based association test

The LD structure of the associated region was determined using HAPLOVIEW and the solid spine of LD algorithm. A total of 14 haplotype blocks, with 64 haplotypes were identified (Fig 3.10) across a 3 Mb region which encompassed the 2.92 Mb FIS associated region. Permuted association tests (10,000 permutations) were carried out on the identified haplotype blocks, identifying 15 haplotype blocks with significant association i.e. with P -values < 0.05 (Table 3.3). The highest association was observed with haplotype block 4 (consisting of two SNPs: BIEC2-692673 and BIEC2-692674), displaying a P -value_(mperm) of 1.0×10^{-4} . This haplotype block had four unique haplotypes, one of which was homozygous in all of the affected animals used in this study. One of these SNPs (BIEC2-692674) was the SNP that consistently gave a statistically significant association in the case-control allelic study. The LD structure was plotted against the associated P -values (case-control and family-based association) in the FIS associated region (Fig 3.10). The amount of LD between the QTL and the marker is specified as D-prime ($0 < d < 1$). A value of 0 indicates that the two loci are in complete equilibrium, whereas 1 represents the highest amount of disequilibrium possible is present (this amount depends of the relative allele frequencies of QTL and marker - i.e. complete disequilibrium could never be observed if the allele frequencies are different at QTL and marker).

Haplotype Block	Permutation p-value	1st SNP in haplotype	SNPs in the computed haplotype
Block 4: GC	0.0001	BIEC2-692673	BIEC2-692673 BIEC2692674
Block 11: GGGGC	0.0003	BIEC2-693015	BIEC2-693015 BIECS693019 BIEC2-693020 BIEC2693028 BIECS693044
Block 12: CC	0.0003	BIEC2-693058	BIEC2-693058 BIEC2-693062
Block 12: TT	0.0003	BIEC2-693058	BIEC2-693058 BIEC2-693062
Block 13: TATTGT	0.0003	BIEC2-693113	BIEC2-693113 BIEC2-693115 BIEC2-693137 BIEC2-693138 BIEC2-693436 BIEC2-693438
Block 3: TT	0.0026	BIEC2-692644	BIEC2-692644 BIEC2692645
Block 3: CG	0.0026	BIEC2-692644	BIEC2-692644 BIEC2692645
Block 5: CCA GT	0.0029	BIEC2-692696	BIEC2-692696 BIEC2-692750 BIEC2692752 BIEC2-692781 BIEC2-692793
Block 4: TA	0.0031	BIEC2-692673	BIEC2-692673 BIEC2692674
Block 11: GGGAC	0.0048	BIEC2-693015	BIEC2-693015 BIECS693019 BIEC2-693020 BIEC2693028 BIECS693044
Block 8: AAC	0.0181	BIEC2-692899	BIEC2-692899 BIEC2-692941 BIEC2-692944
Block 9: CTCC	0.0181	BIEC2-692977	BIEC2-692977 BIEC2-692983 BIEC2-692987 BIEC2-692993
Block 10: CTCC	0.0181	BIEC2-692997	BIEC2-692997 BIEC2-692998 BIEC2-693002 BIEC2-693007
Block 2: ACACCGCG	0.0189	BIEC2-692533	BIEC2-692543 BIEC2-692550 BIEC2-692563 BIEC2-692602 BIEC2-692619 BIEC2-692621 BIEC2-692640
Block 8: GGA	0.0206	BIEC2-692899	BIEC2-692899 BIEC2-692941 BIEC2-692944

Table 3.3: Haplotype blocks identified with significant association to FIS. A total of 64 haplotypes in 14 haplotype blocks were identified using the solid-spine of LD algorithm in HAPLOVIEW. A case-control haplotype association test over 10,000 permutations identified 15 of these haplotypes as displaying significant disease association.

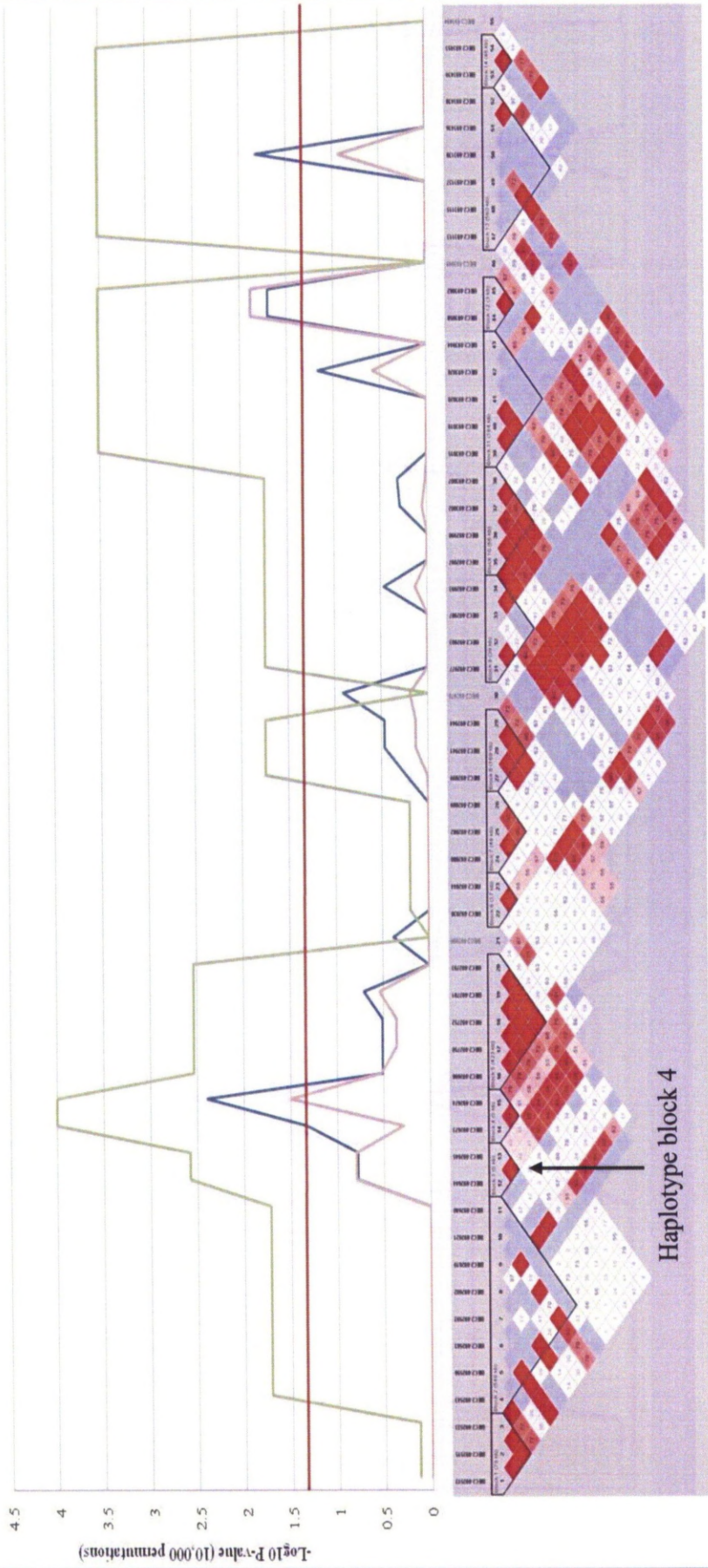


Figure 3.10.

Legend to figure 3.10: *Linkage disequilibrium of a 3 Mb region on ECA26, encompassing the 2.92 Mb region which displays significant FIS association. Plotted at the top are the corresponding $-\log_{10}$ of P for the allelic case-control (blue line), the family-based analyses (pink line) and the highest observed $-\log_{10}$ of P for each haplotype block (green line). The horizontal red line indicates the formal statistical significance (P-value 0.05). Significance for all three tests was calculated over 10,000 permutations. The bottom plot shows LD between SNPs in the corresponding region. Red diamonds represent D' values equal to 1, lower values of D' are shaded from pink to white, while SNPs that are not in LD to one another are shown in blue. Thick lines around SNP groups represent the 14 defined haplotypes.*

3.4. Discussion

The Fell Pony breed was founded approximately 115 years ago with no more than 30 animals. Therefore, like many other domesticated breeds, relatedness within the breed is likely to be high. In addition there have been recent genetic bottlenecks and inbreeding has been common practice. A potential issue with performing case-control association studies using related samples is that it can lead to spurious associations arising from an inflation of the observed P -values. A common founder has been identified for all of the affected samples used in this study. Based on this, relatedness between samples is likely to be high and this should be accounted for. An approach to overcome the risks of performing a case-control study with related samples is to perform a family-based association test. Consequently, both a case-control and family-based association study were performed to identify a reliable disease association. Both sets of analyses led to the identification of the same associated 2.92 Mb (29.69 – 32.18 Mb) region. Further inspection revealed that 13 of the 44 SNPs across the identified region formed a contiguous block of homozygosity in all of the affected animals. This region defines a ~1 Mb region of homozygosity (29,803,727 – 30,802,367) that encompasses the microsatellite marker (NVHEQ070 at 30.25 Mb) which displayed significant FIS-linkage in the microsatellite whole-genome scan (Chapter Two). Although the two approaches did yield similar results, the family-based method did not display statistically significant disease linkage whereas the case-control based method did. The case-control study produced generally higher P -values suggesting that relatedness in samples can lead to an inflation of the observed P -values. P -values for both the case-control and family-based association tests were corrected for multiple testing using 10,000 permutations, again revealing statistically significant disease association. The association was observed for the same 2.92 Mb region, with P -values across the remainder of the genome being >0.71 in the case control studies and >0.28 in the family-based test. The case-control approach gave the lowest P -value of 0.004 at 29.80 Mb and the family-based association identified 31.28 and 31.59 Mb with P -values of 0.042. The level of significance across the 2.92 Mb region differed slightly

between the two approaches, identifying alternative statistically significant loci, but when the results were plotted the profile of the P -values mirrored one-another (Fig 3.8).

Haplotype based studies are extremely useful when mapping diseases in populations where linkage disequilibrium is low, to increase power. In an attempt to further map the associated loci within the identified 2.92 Mb region, LD between SNPs in the region was assessed using HAPLOVIEW and then haplotype blocks identified. Haplotype blocks were inferred by using the ‘solid spine of LD’ algorithm and then a haplotype based case-control association, corrected for multiple testing with 10,000 permutations, was performed. This analysis identified 15 haplotypes showing significant disease linkage, revealing a higher association (P -value) across the critical region then revealed with the case-control and family-based study. The 13 SNPs which were homozygous in all FIS-affected animals, formed four haplotype blocks, of which only two (blocks four and five) showed significant disease association. Haplotype block four which contained the SNP which displayed the highest disease association in the case-control study (BIEC2-692674), also showed the most significant disease linkage (P -value 0.0001) with the haplotype analysis. Furthermore, this haplotype was shared by all 18 affected animals used in this study. Haplotype block 12, a two SNP haplotype of BIEC2-693058 and BIEC2-693062, which was identified by the family-based studies, was also identified as showing a significant association (P -value 0.0003), with further inspection showing that 14 of the affected animals shared this common haplotype.

The aim of this study was to definitively map the chromosomal region where the FIS mutation was most likely to lie. The mapping is unambiguous: the genome-wide P -values for the identified association on chromosome 26, is 54 times more significant than that for any other observed in the entire genome for the family-based study and 177 times more significant in the case-control association test. Statistical analysis identified that the FIS mutation maps to a 2.92 Mb region on chromosome 26. By further examining the haplotypes of the affected animals across this association 2.92 Mb region, a ~1 Mb (29,803,727 – 30,802,367) homozygous IBD haplotype was identified, spanning 13 SNPs in a contiguous block.

This study has demonstrated the value of the EquineSNP50 Beadchip and its use in mapping disease loci where the limited availability of related samples may impede the progress of microsatellite based linkage studies. Here, an association was mapped for a simple trait using relatively few samples and it was demonstrated that case-control GWAS can be successfully applied to mapping disease loci in related animals.

The association study identified a ~1 Mb region of homozygosity in the 13 affected animals used for this experiment, within a 2.92 Mb region which displays significant disease association. Therefore further fine-mapping studies are now required to map precisely the FIS associated haplotype, and the results of this are discussed in the following chapter. This will be performed by using additional genetic markers to identify the smallest homozygous IBD haplotype which is shared by all of the affected animals, and is the haplotype on which the mutation arose. Once the critical region has been defined, it will be examined for genes which are deemed suitable candidates, based on phenotypic similarities in other species and on gene function. These candidate genes will then be sequencing in an attempt to identify the causal variant which is responsible for FIS.

Chapter 4

Fine mapping the critical region and interrogation of candidate genes

	Page
Summary	92
4.1. Introduction	92
Fine-mapping studies to narrow the critical chromosomal region	92
Selection of genetic markers for fine-mapping critical intervals in the horse	93
Selection and interrogation of candidate genes	94
Aims and Objectives of narrowing the critical chromosomal region and interrogation of potential candidate genes	96
4.2 Materials and methods	97
Animals and Samples	97
Fine-structure mapping	98
Examination of candidate genes in an attempt to identify the causal mutation	101
4.3 Results	104
Fine-mapping the homozygous critical region	104
Interrogation of the critical region – searching for candidate genes	107
4.4 Discussion	111

Summary

Mapping studies, including a microsatellite whole-genome scan and a genome-wide association scan, have definitively confirmed that the chromosomal defect responsible for FIS lies on ECA26. To further refine the location of the disease gene within the boundaries provided by the linkage based and association studies, fine-mapping studies were undertaken to map the homozygous IBD haplotype is shared by all of the affected animals investigated in this study. Once the critical interval was definitively defined, positional candidate genes were examined in an attempt to identify the causal mutation, and although polymorphisms were identified the causal variant was not found.

4.1. Introduction

Fine-mapping studies to narrow the critical chromosomal region

After demonstrating linkage and/or an association, further mapping of recombination break-points can usually be used to refine the critical region further, thus reducing the number of genes which require interrogation. This approach examines the critical interval with additional genetic markers to identify recombination events in affected individuals that can be used to further narrow the shared homozygous haplotype. The length of the homozygous haplotype will differ between affected animals, dependent on the position of recombination events which have broken down the ancestral haplotype, but will always contain the same block of homozygosity surrounding the causal variant, representing the ancestral haplotype on which the mutation arose. The existence of LD between the disease allele and nearby markers is commonly used to fine-map the disease interval, an approach which has proved successful in many species (Parker et al., 2007, Peterfy et al., 2006, Bellone et al., 2010), and often leads the investigator to the causal variant.

Fine-mapping with genetic markers is usually performed in two stages: Initially microsatellite markers are used as they are highly polymorphic as a result of their

high mutation rate which is estimated to be $100,000 \times$ greater than single nucleotide polymorphisms (SNPs) (Fondon and Garner, 2004). Secondly, SNPs are used to refine the boundaries, by ‘walking-in’ from the break-points of homozygosity which have been identified by the microsatellite markers. A potential problem when fine-mapping diseases in domesticated species which have undergone genetic bottlenecks resulting from small numbers of founders, over-use of popular sires, and fluctuations in population size, is that there will be reduced heterogeneity and increased average length of LD (Parker et al., 2007). Consequently, further fine-mapping may be impossible, leaving the researcher with an interval which could extend over several megabases, containing many candidate genes and making follow-up studies complex and expensive. An approach to overcome this, which is commonly used in canine studies (Goldstein et al., 2006, Karlsson et al., 2007) is the two-stage approach. This uses long within-breed LD to map the disease linked region and then uses shorter inter-breed LD to refine the interval. Thus, by combining data from multiple breeds the interval can often be refined to a much shorter interval. However, in order to utilise this approach, more than one breed must carry the same disease trait. As described in chapter one, FIS has also been identified in the Dales Pony. Therefore, use of the Dales breed in the mapping study may be useful for defining a smaller haplotype for candidate gene screening, although only a single FIS-affected Dales sample was available for this.

Selection of genetic markers for fine-mapping critical intervals in the horse

As discussed in chapter two, there are over 24,000 microsatellite sequences that have been submitted to the NCBI database. Dependent on the number of microsatellites identified in the chromosomal region of interest from those submitted to NCBI, additional markers may be required to supplement these and provide denser coverage. The horse genome, which became publically available in January 2007, has enabled the identification of microsatellites to become a relatively simple task. To identify additional microsatellite markers, the reference sequence of the target region can be downloaded using the ‘export data’ tool on the ENSEMBL genome browser (http://www.ensembl.org/Equus_caballus/Info/Index) or using the ‘Get DNA tool’ on the UCSC browser (<http://genome.ucsc.edu/cgi->

bin/hgGateway?org=Horse&db=equCab2&hgsid=170601515) and examined for variable number tandem repeats.

Sequencing the horse genome led to the identification of approximately one million SNPs for both genome-wide and fine-mapping studies. All of the SNPs used in this work are described in the Broad Institute database (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/).

Selection and interrogation of candidate genes

Once the critical interval has been narrowed as much as possible, genes in the interval should be examined for likely candidates that are worthy of further interrogation. All genes within a critical interval remain potential candidate genes; however, examining the function of the genes may enable prioritisation. The conventional and more commonly used approach to identify potential candidates usually involves searching for literature on functional annotation. Additionally, examining for phenotypic similarities in targeted knockout mice and homologous diseases in man or other species (Fig. 4.1) using databases such as Online Mendelian Inheritance in Man (OMIM) or Online Mendelian Inheritance in Animals (OMIA) can help to identify potential candidate genes. More recently, automated and interactive approaches have been developed to help the researcher select candidate genes. They offer an alternative approach, enabling the researcher to prioritise which genes to examine, without the need to investigate the functions of each gene individually. However, the algorithms used by these programs to select candidate genes, are largely inaccessible (Seelow et al., 2008). Furthermore, many of these programs are designed specifically for the human genome, making them unsuitable for animal studies. Therefore, the traditional approach of searching databases and reviewing literature remains favorable, particularly in animal studies and when dealing with a manageably sized critical interval harboring relatively few genes.

When the final selection of potential candidate genes have been identified, the genes would then be sequenced, with the coding regions as a priority, to identify a novel mutation which has a genotypic pattern that corresponds to that predicted for the inheritance model under investigation.

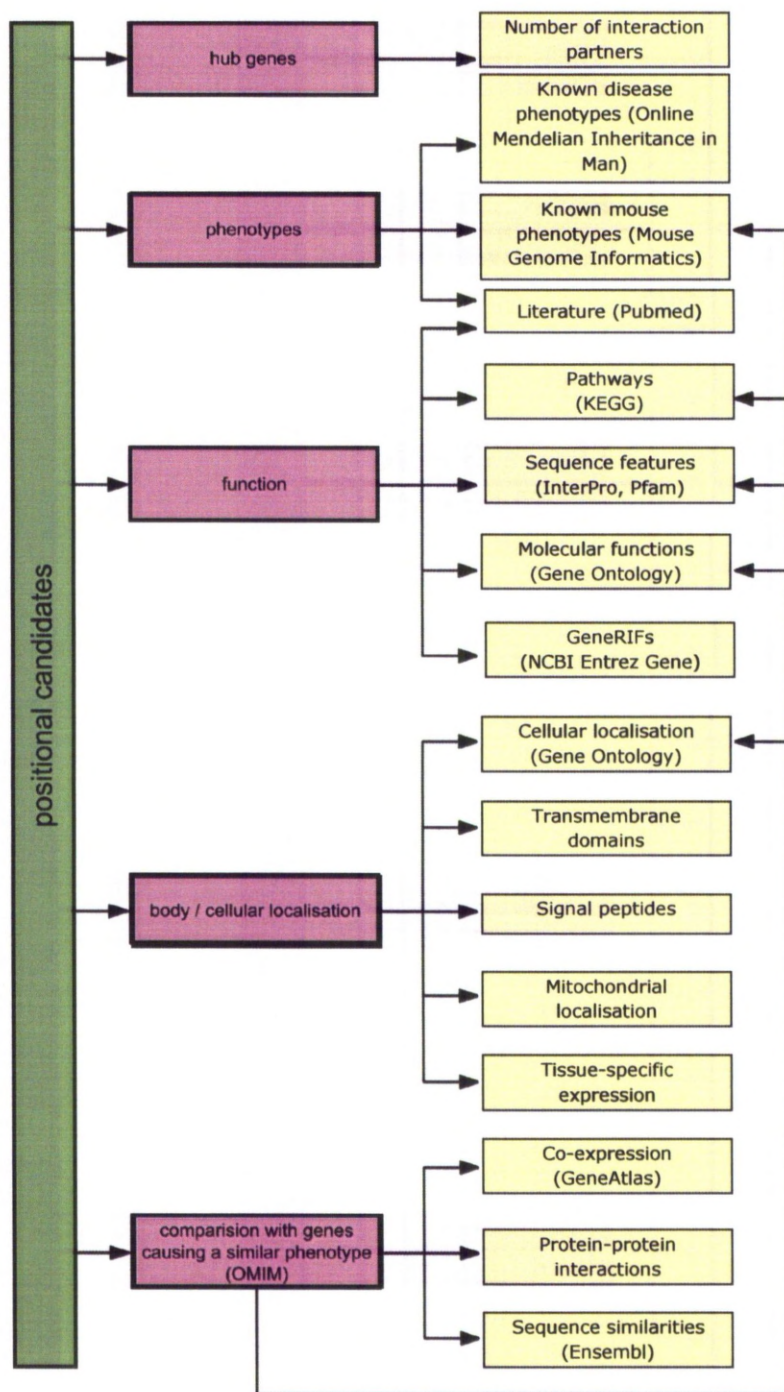


Fig. 4.1. Illustration of the main approaches used for selecting candidate genes from a linkage interval. The general concepts are depicted as pink boxes, gene properties are shown in the yellow boxes with the database for the query shown in brackets. Illustration adapted from Seelow et al, 2008.

Aims and Objectives of narrowing the critical chromosomal region and interrogation of potential candidate genes

Fine-mapping and candidate gene screening has proven a successful approach for identifying disease genes in domesticated species (Patterson et al., 2008, Tryon et al., 2007). Based on this, the interval described in chapters two and three, that showed significant disease association, was further refined by screening additional genetic markers in FIS-affected samples to identify the homozygous IBD haplotype which encompasses the FIS mutation. When the interval had been refined as much as possible using genetic markers, it was then examined for potential candidate genes using the traditional method of reviewing literature, functional annotation and phenotypic similarities in other species. To further complement this, freely available web-based software (GeneDistiller), for automatic identification of potential candidates was also used, enabling a comparison of the traditional and newly developed methods. The coding sequence of the identified potential candidate genes was then sequenced to search for possible causal variants.

The aims of the fine-mapping and candidate gene studies were to:

- a) Map definitively the boundaries of the homozygous haplotype shared by all twelve of the FIS-affected Fell Pony samples used in this study, using both microsatellites and SNPs.
- b) In an attempt to further refine the homozygous haplotype, the extent of the shared haplotype would be examined in the FIS-affected Dales sample.
- c) Examine the critical interval for candidate genes, sequence these, and identify possible causal variants for further investigation.

4.2 Materials and methods

Animals and Samples

A panel of nine obligate carriers were initially used to screen markers to identify if they were polymorphic in the Fell Pony breed. Twelve affected Fell Pony samples, collected by Dr. G Thomas (Thomas, 2003) were chosen for fine-mapping (figure 4.2) studies to identify the boundaries of the FIS critical interval. Additionally, sample 03_08 was selected for fine-mapping, as it had the smallest observed homozygous haplotype identified from the SNP GWAS. These 22 samples (appendix 1) had already had DNA extracted from them for use in the microsatellite whole-genome scan and the genome-wide association study (GWAS) (for details on the sample collection and extraction methods, see chapters 2.2 and 3.2). Nine carriers and nine of the affected samples were used in the GWAS study; the remaining four affected samples were used in the microsatellite scan. In addition, a single Dales affected sample was used (see next section). All DNA samples were quantified using a Nanodrop® and an aliquot of stock DNA diluted to an approximate concentration of 10ng/μL using ddH₂O.

Sample collection from the first confirmed FIS-affected Dales foal

Jugular blood and tissue samples were collected during the post-mortem examination at the University of Liverpool, by the author and F. Malalana (Equine Division, Dept Veterinary Clinical Science), immediately after euthanasia of the animal (Fox-Clipsham et al., 2009)). Samples were collected with prior consent from the owner. Jugular blood samples were collected into Vacutainers® containing the anticoagulant ethylenediamine tetra-acetic acid (EDTA) (Beckton Dickinson, UK) before archiving at -20°C. Tissue samples (kidney, spleen, muscle, liver, bone marrow and lymph nodes) were collected during post-mortem examination for histological and immunohistochemical analysis. These samples were archived as fresh tissues at -20°C.

DNA extraction from the Dales foal sample

DNA was isolated from kidney using a revised version of the Nucleon® blood extraction kit protocol (Appendix 5; GE Healthcare, UK). DNA was extracted from 150mg of kidney tissue and re-suspended in TE buffer (Sigma-Aldrich, UK). The concentration of the DNA was assessed using a Nanodrop® by measuring absorbance at 260nm. An aliquot of the stock DNA was further diluted to an approximate concentration of 10ng/μL using ddH₂O.

Fine-structure mapping

Initially, microsatellite markers that were near the NVHEQ070 marker were used to further refine the interval. Following this, SNPs were used to fine map more precisely the boundaries of the homozygous IBD haplotype shared by all 12 affected animals. To assess informativeness of the genetic markers, they were first analysed in the obligate carrier samples to ensure that they were polymorphic in the Fell Pony breed.

Designing primers for amplifying a fragment which contains a genetic marker

Primers for interrogating microsatellites and SNPs were designed using Primer3 v4.0 online software (<http://frodo.wi.mit.edu/primer3/>). The target sequence was imported into Primer3, with at least 100 bp flanking sequence at the 5' and 3' ends, and primers designed no less than 50 bp from either end of the target sequence. The maximum amplicon length was limited to 500bp for those containing microsatellites and 800bp for the sequencing of amplicons containing SNPs. Primer lengths of >18 bp and <27 bp (optimum size of 21 bp) were stipulated and a GC content >40% and <60%. Finally, an 18 bp universal tail (TGACCGGCAGCAAATTG) was added to the forward primer and primers ordered from Sigma-Aldrich (desalted with a standardised concentration of 100μM).

Refining the critical region using microsatellite markers

The SNP GWAS had identified significant association to a ~2.92 Mb interval; however, closer inspection of the whole-genome microsatellite scan data revealed that marker TKY1155 (ECA26; 29,807,492bp) was heterozygous in multiple affected animals. Therefore, only markers downstream of this were selected for fine-mapping the homozygous haplotype. The associated interval for fine-mapping was thereby reduced to ~2.38 Mb, from marker TKY1155 at 29,807,492 bp to BIEC2-693138 at 32,187,972. Three microsatellite markers (appendix 11) were identified for fine-mapping from those submitted to the National Centre for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/nuccore>). The physical position of the three markers was confirmed on the horse genome assembly EquCab2.0 using the BLAST-like alignment tool (BLAT) in ENSEMBL (<http://www.ensembl.org/Multi/blastview>). Markers where multiple alignments (did not map uniquely) were observed were excluded from further analysis. In an attempt to identify further microsatellite markers in the ~2.38 Mb region the reference sequence was downloaded and examined for additional di-nucleotide microsatellite markers.

Genotyping of the microsatellite markers was performed in a PCR; initially the nine obligate carriers were used to confirm polymorphism in the breed. Amplification for fragment analysis was performed in 96-well skirted PCR plates (Axygen, USA) in 12µl reaction volumes, using 20ng gDNA, 0.75 unit AmpliTaq Gold (Applied Biosystems, Foster City, CA), 1 × GeneAmp PCR buffer II (Applied Biosystems, USA), 1.5mM MgCl₂, and 200µM each dNTP. Then 2.5 pmol of reverse, 1 pmol of tailed-forward, and 5 pmol of the labelled universal primer (6-FAM) were added to the reaction. A PCR program of 94°C for 10 min, followed by 30 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by 8 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 1 min, and then 72°C for 30 min was performed using an MJ Tetrad PCR cycler (Bio-Rad Laboratories, Hercules, CA). A 3µl aliquot of the PCR reaction was loaded onto the ABI3100 (Applied Biosystems, USA) for analysis according to the manufacturer's instructions. Dye set G5 was used in conjunction with the LIZ500 size standard.

Genotyping data was collected and analysed using GeneMapper version 4.0 (Applied Biosystems, USA). Allele sizes were assigned to pre-defined bins and automatically given an appropriate integer value. Genotypes were verified by eye to check for genotyping errors. The data was then exported into an Excel spreadsheet and polymorphic markers identified, which were then genotyped in the 12 affected animals following the same protocol. The genotyping data for the affected animals was then visually inspected for heterozygosity, which was used to narrow the shared haplotype.

Refining the critical region using single nucleotide polymorphisms

A total of 118 SNPs were identified across the homozygous region which had been defined by the microsatellite fine mapping. SNPs were selected from those publically available online from the Broad Institute (http://www.broad.mit.edu/ftp/distribution/horse_snp_release/v2/). Initially, the SNPs were sequenced in the nine obligate carrier samples to identify if the SNPs were polymorphic in the Fell Pony breed. Once this had been confirmed, all polymorphic SNPs were sequenced in the 12 affected animals to further refine the homozygous region.

Amplification of the target sequence containing the informative SNP were performed by PCR in a 96-well skirted PCR plate (Axygen, USA) in 12µl volumes, using 20ng gDNA, 0.75 unit AmpliTaq Gold (Applied Biosystems, Foster City, CA), 1 × GeneAmp PCR buffer II (Applied Biosystems, USA), 1.5mM MgCl₂, and 200µM each dNTP; 10 pmol of reverse, 10 pmol of tailed-forward primer were then added to the reaction. Samples were amplified using an MJ Tetrad PCR cycler (Bio-Rad Laboratories, Hercules, CA). A PCR program of 94°C for 10 min, followed by 30 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 2 min, and then 72°C for 10 min was used. The PCR product was then purified using MultiScreen PCR₉₆ 96-well filter plates (Millipore®, USA). PCR products were diluted in 200ul of ddH₂O and then transferred to the wells of the Multiscreen plate. The plate was placed on a vacuum manifold at 24inch mercury for 20 min. Following this, 20µl of ddH₂O was added to the individual wells and incubated at 37°C for 10 min on a plate tilter and then transferred a 96-well skirted PCR plate (Axygen, USA) for storage at -20°C. The sequencing reaction was performed in 96-well PCR plates (Applied Biosystems,

USA) in a 6µl volume using 0.5µl of 5 × BigDye Terminator v3.1 (Applied Biosystems, USA), 1µl of PCR template (5–20 ng), 1µl of 1× BigDye sequencing buffer (Applied Biosystems, USA) and 3.2pmol universal sequencing primer (5'-TGACCGGCAGCAAATTG -3') (Sigma-Aldrich). In addition, for amplicon sizes of >500 bp, the target sequence was also sequenced in the reverse direction, using the reverse primer in place of the universal sequencing primer. A PCR program of 96°C for 0.5 min, followed by 44 cycles of 92°C for 4 sec, 58°C for 4 sec and 72°C for 1.5 min was performed using an MJ Tetrad PCR cycler (Bio-Rad Laboratories, Hercules, CA). To remove unincorporated dye terminators prior to electrophoresis, the sequencing product was purified by precipitation: 60µl of 80% isopropanol (Fisher Scientific, UK) was added to the wells and the plate spun for 30 min at 4000 rpm, the supernatant was then discarded and 100µl of 60% isopropanol (Fisher Scientific, UK) was added to the wells and the plate spun for 10 min at 4000 rpm and the supernatant discarded, followed by a further spin with the plate upside down for 1 min at 1000 rpm. The plate was then left to dry at room temperature for 10 min before re-suspending the product in 10µl of deionized formamide (Applied Biosystems, USA) and loading onto an ABI3100 (Applied Biosystems, USA) for analysis according to the manufacturer's instructions.

Sequencing traces were then visualised using STADEN software (<http://staden.sourceforge.net/>) to identify polymorphic markers; the 12 affected animals were then subsequently genotyped with these markers. The genotyping data was then visually inspected for heterozygosity, which was used to narrow the shared haplotype.

Examination of candidate genes in an attempt to identify the causal mutation

The genotyping data from the whole-genome microsatellite scan, the SNP-based GWAS and the fine-mapping studies using microsatellites and SNPs, was compiled to identify the definitive FIS homozygous haplotype. This homozygous critical region was then examined for potential candidate genes. These genes were selected from those predicted in the ENSEMBL browser (http://www.ensembl.org/Equus_caballus/Info/Index) and the UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>). Candidate genes were

determined based on the function of the gene, phenotypic similarities between FIS and diseases which had been observed in other species, reviewing the Genecards (<http://www.genecards.org/>) and, where knockout mice had been made, reviewing the observed phenotype (<http://www.informatics.jax.org/>). To complement this selection, a web based program, GeneDistiller (<http://www.genedistiller.org/>) (Seelow et al., 2008) was also used to identify potential candidates. This program is designed specifically for human based studies as a complementary program to subsidise the information identified from traditional candidate gene screening. Therefore, before analysis can be performed, the corresponding human chromosomal region is identified. The corresponding human chromosomal region is identified by using the 'Synteny' option in ENSEMBL. Although it is possible that rearrangements may have occurred, the corresponding human region is examined to confirm that all of the expected genes are present. Key phenotypes of FIS are anaemia, b-lymphocyte deficiency and opportunistic infections. Based on the results obtained, a subset of those candidates identified through conventional methods and from the automated detection program is taken forward for mutation screening.

Following candidate gene selection, the predictions made by ENSEMBL and the UCSC genome browser for the equine genome were manually inspected and compared to the human genome for possible prediction errors. Where fewer exons were predicted in the equine genome than observed in the human genome, the additional exonic were mapped in the equine genome using the BLAT tool in ENSEMBL (<http://www.ensembl.org/Multi/blastview>). After confirming the exon-intron boundaries, primers were designed to amplify all of the exons of the candidate genes, including intronic flanking sequence to that the intron-exon boundaries could also be examined.

Sequencing candidate genes

The target sequence incorporating the exon was amplified using the same protocol as used to amplify the SNPs in the fine-mapping studies (shown on page 99). Sequencing traces were then visualised using STADEN software (<http://staden.sourceforge.net/>) to identify heterozygous variants in the obligate carriers. Sequencing was initially performed in the nine obligate carrier samples, to

identify any heterozygous variants suitable for subsequent screening in affected samples.

4.3 Results

Fine-mapping the homozygous critical region

Of the three microsatellite markers that were identified for fine-mapping, from those publically available on the NCBI database (TKY2012, TKY3045 and TKY3044), two were excluded as they did not map uniquely (mapped to more than one region) within the equine genome (TKY3045 and TKY3044). Four further potential genetic markers were identified by examining the critical interval for di-nucleotide sequences; these were named FPS1, FPS2, FPS3 and FPS4. Obligate carriers were then screened with the five microsatellites to identify if they were polymorphic; one of the microsatellites (FPS4) was monomorphic so excluded from further analysis. The remaining four microsatellites were then used to genotype the 12 affected Fell Pony samples. To complete the dataset, NVHEQ070 and TKY1155, which were genotyped as part of the whole-genome scan, were also genotyped with this sample set. Of the 12 affected animals screened with the seven markers, all 12 showed homozygosity centred on the NVHEQ070 marker which was used in the initial whole-genome microsatellite scan. Two animals were homozygous for all six markers, a region which spans 2.64 Mb. The remaining 11 animals contained a smaller block of homozygosity, proximal and/or distal to but always containing four of the markers; FPS1, FPS2, FPS3 and NVHEQ070 (Fig. 4.2.). Further fine-mapping with SNPs would now enable more precise and accurate fine-mapping of the critical interval.

Of the 118 SNPs screened, 62 were polymorphic (appendix 11) in the obligate carriers and were subsequently used to screen the 12 affected animals. In addition, a further 15 novel SNPs were identified from the sequencing traces. Given this, a total of 77 SNPs were examined in the affected animals (appendix 6). Of the 12 affected animals, five were homozygous for all 77 SNPs so they did not narrow the haplotype further than that already achieved by the microsatellite data. The smallest homozygous haplotype spanning 992,357 bp (29,825,158 bp – 30,817,514 bp) was defined by four individuals; 09_00 (A3), 97_00 (A5), 43_02 (A6) and 03_08 (A13).

The SNP which defined the boundary of the critical interval at the 3'-end was the same SNP (BIEC2-692880) which had been identified from the SNP GWAS. This SNP genotyping was confirmed using Sanger sequencing.

Now that the critical region had been defined as much as possible using the Fell Pony FIS-affected samples, the Dales FIS-affected foal was screened for genetic markers in an attempt to further refine and narrow this critical interval in this breed. Those SNPs in the 992,356 bp critical interval were genotyped in this animal. All 11 SNPs were homozygous, providing no further refinement to the region.

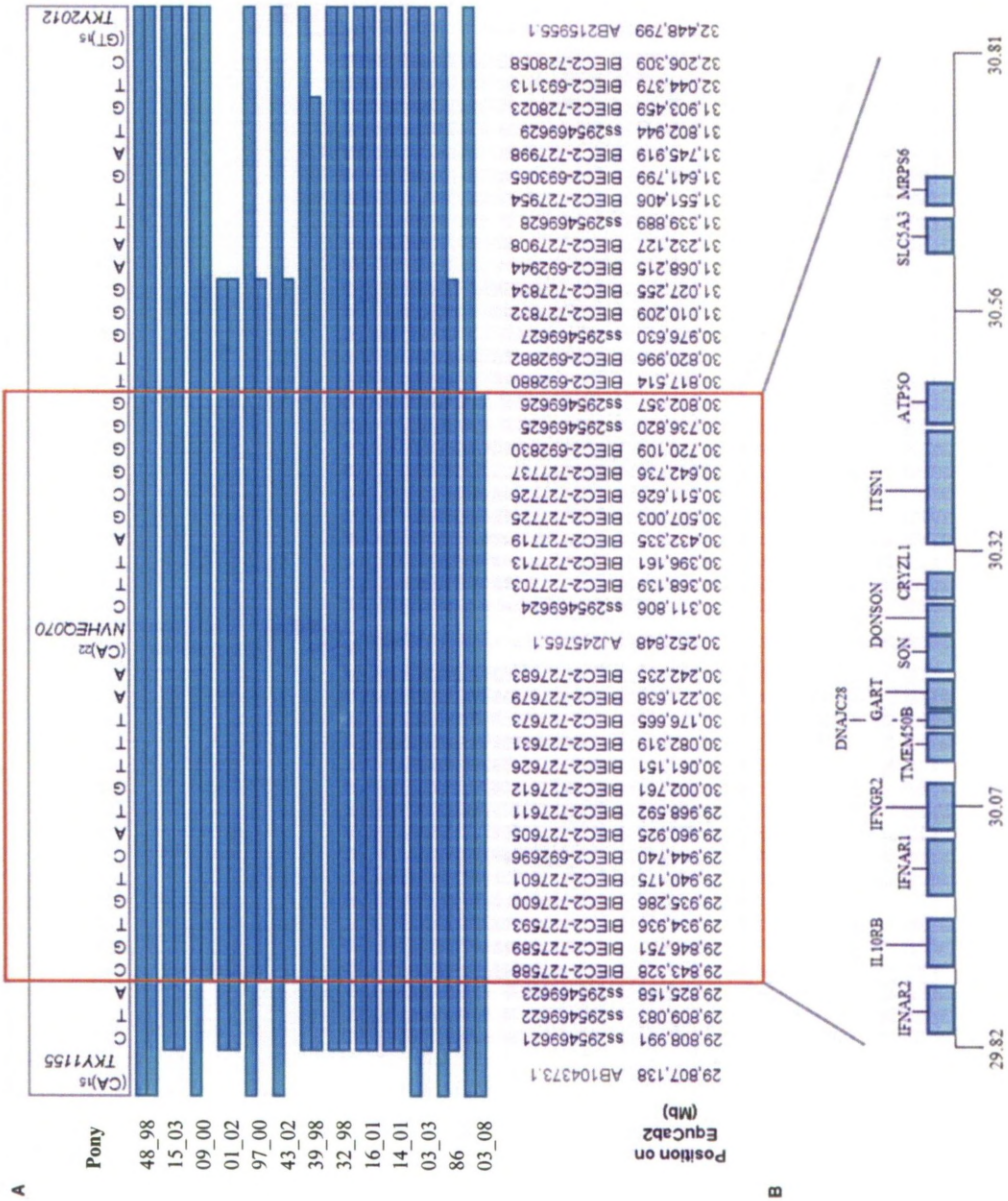


Figure 4.2: The affected SNP haplotype is shown at the top. The affected alleles for three key microsatellite markers are also included. The extent of conserved affected haplotypes present in the 13 affected individuals (A1-13) are indicated with blue bars. Accession numbers (newly identified SNPs) or the local SNP ID number (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/), together with the genome position are given. The minimal 992 kb shared region of homozygosity (29,825 – 30,817 kb) is out-lined in orange. (b) The positions of the 13 ENSEMBL annotated genes within the conserved block are indicated.

Interrogation of the critical region – searching for candidate genes

Thirteen predicted genes lay in the 992,356 bp critical region (Fig. 4.3). Of these, four were deemed good candidates based on function and, where possible, phenotypic similarities in knockout mice. These were *IFNAR2*, *IL10RB*, *IFNAR1* and *IFNGR2* - four functionally related genes which form a cytokine receptor gene cluster on chromosome 26.

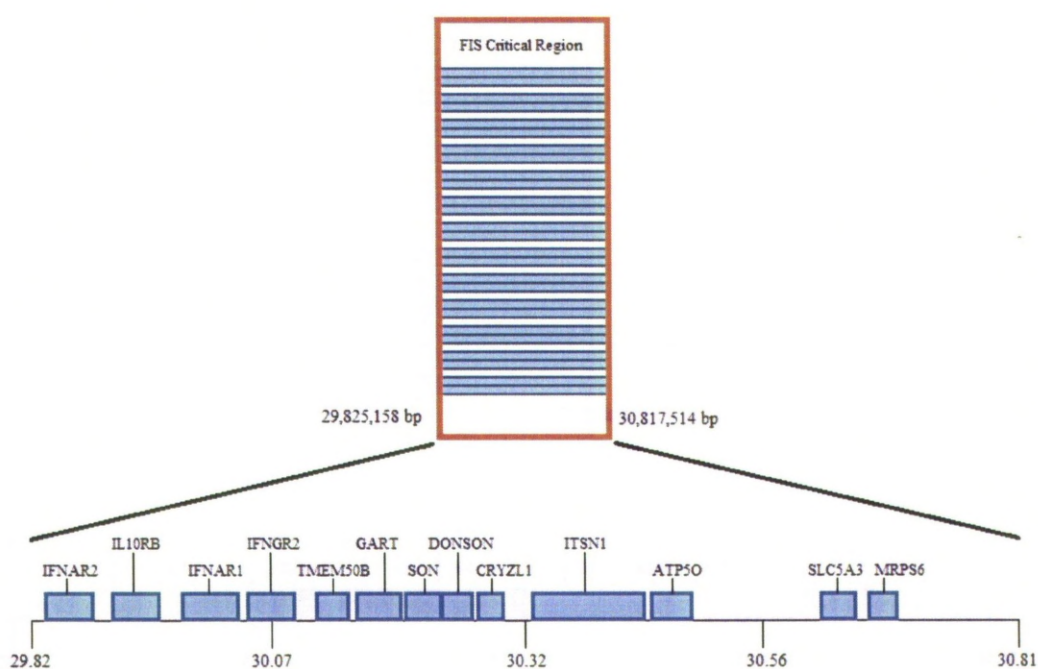


Fig. 4.3. Identification of annotated genes: Position (Mb) of the 13 genes which are annotated on ENSEMBL and fall within the 992,356 bp FIS homozygous critical region.

The corresponding human chromosomal region was then identified from ENSEMBL by searching for genes which had been identified in the critical interval; this was HSA 21: 34.59 Mb – 35.48 Mb. Although some slight rearrangement of the genes was observed, all 13 were identified in the corresponding human region. GeneDistiller identified four potential candidates: *IFNAR1*, *IL10RB*, *IFNAR2* and *IFNGR2*.

As both methods had identified the same candidate genes, all four were selected for sequencing analysis. Exons of the four genes were sequenced, along with >50 bp of flanking intronic sequence so that exon-intron boundaries could also be examined for mutations.

Sequencing the four candidate genes

The exons of the four genes were sequenced in nine obligate carriers, identifying multiple variants (Table 4.1). An autosomal recessive defect is responsible for FIS and so the causal variant would be heterozygous in all of the obligate carriers.

The interferon (alpha, beta and omega) receptor 2 (*IFNAR2*) and the interferon (alpha, beta and omega) receptor 1 (*IFNAR1*) are essential for mediating type I interferon responses in both haematopoiesis and in the immunological response to infection (de Weerd et al., 2007). Furthermore, mice with targeted mutations or knockouts of these genes have shown severe disruption to the following systems: immune (abnormal T-lymphocyte physiology and increased viral susceptibility), haematopoiesis (anaemia and enlarged spleen) and abnormal blood homeostasis. A total of three variants were identified from sequencing the 11 exons and exon-intron boundaries of *IFNAR1*; two intronic SNPs and one SNP within the coding sequence of exon 10. This mutation, an A>G SNP, is predicted to cause a missense mutation causing an amino acid substitution from Arginine to Lysine. All three of the mutations were homozygous in six of the nine carriers and therefore based on this cannot be the causal variant. The coding sequence of *IFNAR1* was therefore excluded from any further analysis. Sequencing of the six exons and the exon-intron boundaries of *IFNAR2* identified two intronic novel SNPs, but neither were deemed candidate mutations as neither were heterozygous in all nine obligate carrier samples. The coding sequence of this gene was therefore excluded from further examination.

The signal transduction of IL-10 via its receptor, interleukin 10 receptor, beta (*IL10RB*) can affect a variety of immune functions including inhibition of pro-inflammatory cytokine synthesis and regulation of the growth and function of B-lymphocytes and antigen presenting cells. *IL10RB* mutations in humans have been associated with early on-set inflammatory bowel disease (Glocker et al., 2009) and susceptibility to hepatitis B virus infection (Gong et al., 2009). Furthermore,

knockout and targeted knockout mice have reduced life span, suffer colitis, reduced blood haemoglobin content and increased circulating leukocyte numbers. Sequencing of the six exons of *IL10RB* identified three intronic SNPs and a G>A SNP in exon six that is predicted to cause a synonymous (silent) mutation. Homozygous obligate carriers were observed for all four mutations which were identified in *IL10RB* and therefore could be excluded as potential candidate mutations.

The interferon gamma receptor 2 (*IFNGR2*) is a regulator of the JAK (Janus tyrosine Kinase) STAT (Signal Transducer and Activator of Transcription) signalling cascade, controlling cell proliferation and haematopoiesis. Mutations in *IFNGR2* have been shown to cause Mendelian susceptibility to mycobacterial diseases (MSMD) (Al-Muhsen and Casanova, 2008), with knockout mice developing normally but showing disruption to the immune and the hematopoietic systems. *IFNGR2* in the human has two transcripts, one with eight exons and the other with seven exons. ENSEMBL predicts that the equine *IFNGR2* has only six exons which align with the last six exons of both human transcripts. Using the distant homologies in the BLAST tool on the ENSEMBL website, these two additional exons were not identified in the equine genome. Therefore, only the six predicted exons were sequenced, along with flanking sequences so the exon-intron boundaries could be examined. This led to the identification of four intronic variants and one exonic variant. The intronic SNPs resulted in G>T, T>G, C>A and G>A base changes but homozygotes for the alternative allele were observed in five of the obligate carriers so excluded from further analysis. The exonic variant, A>G SNP in exon four was predicted to cause a synonymous (silent) mutation and excluded from further analysis on the basis that homozygotes for the alternative allele were observed in five of the obligate carriers.

To provide further evidence that the variants identified in the four candidate genes were not the FIS causal variant, all variants were genotyped in 12 affected animals (table 4.1) and all 12 samples were homozygous for the alternative allele.

Gene	Position (bp)	Intronic/Exonic	Obligate carrier genotypes			FIS-affected genotypes		
			Homozygous wild-type	Heterozygous	Homozygous alternative allele	Homozygous wild-type	Heterozygous	Homozygous alternative allele
IFNAR2	29,898,126	Intronic	0	4	5	0	0	12
IFNAR2	29,900,728	Intronic	0	5	4	0	0	12
IL10RB	29,937,396	Intronic	0	1	8	0	0	12
IL10RB	29,937,412	Intronic	0	2	7	0	0	12
IL10RB	29,937,418	Intronic	0	1	8	0	0	12
IL10RB	29,940,956	Exon 7	0	4	5	0	0	12
IFNAR1	29,982,712	Intronic	0	3	6	0	0	12
IFNAR1	29,989,752	Intronic	0	4	5	0	0	12
IFNAR1	29,997,461	Exon 10	0	3	6	0	0	12
IFNGR2	30,074,980	Intronic	0	4	5	0	0	12
IFNGR2	30,074,999	Intronic	0	4	5	0	0	12
IFNGR2	30,078,657	Exon 4	0	4	5	0	0	12
IFNGR2	30,082,319	Intronic	0	4	5	0	0	12
IFNGR2	30,082,329	Intronic	0	4	5	0	0	12

Table 4.1: Variants identified from sequencing four candidate genes in the FIS critical interval.

4.4 Discussion

The primary aim of this study was to confirm the boundaries of the homozygous IBD haplotype that is shared by all affected animals and assumed to be inherited from the common ancestor which founded the FIS mutation. The secondary aim of this study was to identify candidate genes within this homozygous block for subsequent mutation screening for the causal variant.

The results of this study, using 13 affected animals, defined the critical interval to a 992,356 bp region on chromosome 26 (29,825,158 bp – 30,817,514 bp). This region encompasses the original microsatellite marker, NVHEQ070, which displayed statistically significant disease linkage in the whole-genome scan, and the SNP BIEC2-692674, which gave the highest disease association in the case-control GWAS. Further, it has led to the exclusion of the most associated region identified in the family-based association study (~31.589 Mb – 31.593 Mb).

Canine studies now commonly use multiple breeds with the same phenotype to further refine a critical interval, an approach which could have been useful in this study. However, examination of the Dales affected foal's haplotype, using a selection of SNPs spanning the defined 992,356 bp interval, revealed that the Dales foal had an extended track of homozygosity so did not further refine the region. FIS is likely to be a relatively recent mutation which happened after the founding of the Fell breed and therefore, relatively few recombinations are likely to have occurred, breaking down the ancestral haplotype.

As the chromosomal location of the FIS mutation was considerably narrowed, positional candidate gene screening was performed. Both a traditional method of reviewing literature and phenotypic similarities in other species, and an automated computational approach was taken for selecting candidates for mutation screening. Both of these methods identified the same four genes as plausible candidates. These were screened for mutations. A total of 14 novel mutations, of which three were exonic, were identified from sequencing the coding sequence of the four candidate genes along with >50 bp of intronic flanking sequence. However, all were excluded

as causal variants; homozygotes for the alternative allele were observed in multiple obligate carriers, and are therefore not consistent with the genotypic pattern that would be observed in a recessive disease.

Although this study did not yield a potential causal variant in one of the short-listed candidate genes, it has successfully defined the critical interval where the FIS mutation lies. Further studies are now required to interrogate the non-coding sequence of the candidate genes and examine the remaining genes within the critical interval. Manual sequencing of large genomic regions can be both time-consuming and expensive. Consequently, the use of high-throughput re-sequencing seems like a logical next step, as it will enable interrogation of the entire critical region. This approach has been hugely successful in identifying causal mutations in the human (Vermeer et al., 2010, Kahrizi et al., 2011, Nikopoulos et al., 2010). Therefore, further studies will now be undertaken to re-sequence the critical region using a high-throughput re-sequencing platform, and the results of this will be discussed in the following chapter.

Chapter 5

Re-sequencing the critical interval to identify the causal variant

	Page
Summary	113
5.1. Introduction	113
Recent advances in genetics: The impact of next-generation sequencing	113
Sequencing using the Roche FLX Titanium series: Pyrosequencing technology	116
Selective re-sequencing: Capturing the target region	117
Analysis of next-generation sequencing data	118
Objectives of re-sequencing the FIS critical interval to identify the causal mutation	122
5.2. Materials and Methods	123
Animals and Samples	123
DNA extraction from tissue and blood samples	123
Designing the sequence capture array	124
Capturing the target sequence and sequencing the libraries	124
Analysis of the sequencing data	127
5.3. Results	133
Sequence capture, sequencing and mapping the reads	133
Refining the homozygous critical haplotype	134
Investigating large scale rearrangements	140
Identification and interrogation of sequence variants in the critical region	142
5.4. Discussion	145

Summary

In previous chapters, genome-wide mapping and haplotype mapping studies have identified a 992,356 bp region on chromosome 26 as the location of the FIS mutation. Candidate gene sequencing did not identify the causal variant and therefore interrogation of the entire interval was deemed necessary. This chapter will introduce next-generation sequencing and its impact on mapping disease mutations, and will present the results of mapping the FIS mutation using these next-generation sequencing approaches.

5.1. Introduction

Recent advances in genetics: The impact of next-generation sequencing

Chain-termination sequencing, termed Sanger sequencing, was first described in 1977 (Sanger et al., 1977) and to date is still one of the most commonly used methods in genetic research due to its relatively low error rate and long read length (up to 900 bp). However, the use of Sanger sequencing in large sequencing projects has limited use due to high costs and limited throughput. Over recent years new methodologies, termed ‘next-generation sequencing’ have been developed, offering a more cost-effective approach to high-throughput sequencing (Morozova and Marra, 2008). The first commercially available next-generation sequencer was introduced in 2004 by Roche (<http://www.roche.com/index.htm>), and then in 2006 Illumina launched their first next-generation sequencer analyser (<http://www.illumina.com/>), followed by Applied Biosystems in 2007 (<http://www.appliedbiosystems.com/absite/us/en/home.html>).

Next-generation instruments are capable of sequencing millions of reads in parallel, dramatically increasing sequence throughput and decreasing associated costs. In contrast to capillary based sequencing, next generation sequencing reads are produced from fragment ‘libraries’, with preparation of sequencing libraries being fairly straightforward: Fragmentation of the gDNA is followed by the ligation of

specific adaptor oligos to each end of the DNA fragments. The approach to sequencing the libraries differs significantly between the next-generation sequencing platforms, with each technology having its own advantages and limitations (see Table 5.1 for a comparison of the current specifications). Pyrosequencing remains the most expensive of the three systems but due to the relatively long read lengths, it is still considered to be the leading platform (Fox et al., 2009).

Company name and platform	Approach	Read length (bp)	Bp per day	Run time	Advantages	Disadvantages
Roche FLX Titanium Series	Pyrosequencing	Up to 400 bp (mean 250 bp)	1 Gb	1 day	<ul style="list-style-type: none"> • Accuracy >99.5% • Supports the use of NimbleGen capture arrays • Substitution errors are extremely rare • Custom made analysis software 	<ul style="list-style-type: none"> • Error prone from incorrectly estimating the length of homopolymers
Illumina Genome Analyzer	Sequencing-by-synthesis with reversible terminators	35 -150	1.7 Gb	Up to 14 days	<ul style="list-style-type: none"> • Low DNA input <1µg • Reliable homopolymer sequencing 	<ul style="list-style-type: none"> • Lowest accuracy at 98.5% • Substitution errors
Applied Biosystems SOLid	Sequencing by ligation	35 - 75	1 - 3 Gb	Up to 14 days	<ul style="list-style-type: none"> • Each position is probed twice give the highest base-calling accuracy >99.9% 	<ul style="list-style-type: none"> • Short read lengths • Limited published data

Table 5.1: Summary of the three most commonly used next-generation sequencing platforms. Current specifications of the platforms are listed along with the advantages and limitations of these methodologies.

Sequencing using the Roche FLX Titanium series: Pyrosequencing technology

The Roche 454 sequencing system has been used to successfully map many disease causing mutations (Browne, 2010, Nikopoulos et al., 2010) and was consequently chosen for sequencing the critical interval identified in our study, based on its read length, high accuracy, and the availability of custom made analysis software developed specifically for the Roche platform.

The technology used by Roche 454 FLX Titanium series is pyrosequencing, which is a DNA sequencing technology based on the sequencing-by-synthesis principle. The sequencing-by-synthesis principle was first described in the mid-eighties (Melamede, 1985), based on the addition of nucleotides to a primed template, deducing the template sequence from the order in which fluorescently labelled nucleotides are incorporated into the growing template. Pyrosequencing utilises this methodology but rather than using fluorescence, light signalling is used to deduce the incorporated nucleotide (Fig 5.1). The intensity of the light signal is used to correlate the number of nucleotides that have been incorporated, which proves problematic for sequencing homopolymeric regions that are greater than three bases in length (Hert et al., 2008). An average read length of ~250 bp with accuracy >99.5% is typically observed with the Roche FLX Titanium series. Substitution errors are extremely rare ($< 10^{-6}$) with the majority of errors arising from undercalling or overcalling the length of homopolymers (Droege and Hill, 2008).

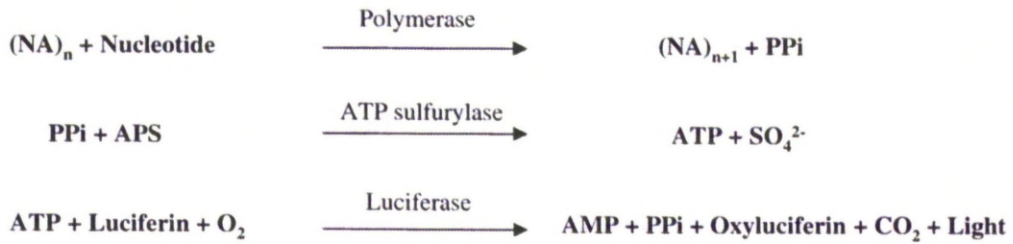


Figure 5.1: *Illustration of the basic principles of pyrosequencing: 1) A complementary nucleotide is incorporated into the growing DNA strand, releasing pyrophosphate. 2) The pyrophosphate acts as a substrate for the enzyme ATP sulfurylase, producing ATP. 3) When the ATP comes into contact with the enzyme luciferase, a light signal is produced that is recorded by the instrument. (Ronaghi, 2001).*

A single run with the Roche FLX Titanium series yields more than 400,000 reads and the platform is capable of performing up to two runs in one day. For smaller projects, where less sequence data is required per sample, sample-specific multiplex identifiers (MIDs) can be used to enable pooling of samples within the same sequencing run and thus increasing efficiency. MIDs are short nucleotide adaptors which are added to the ends of the DNA fragments prior to sequencing; after sequencing they are used to identify the reads for each individual sample.

Selective re-sequencing: Capturing the target region

Traditionally, PCR-based procedures have been used to capture a chromosome target region for selective re-sequencing, but this method is now out-dated because, as next-generation sequencing costs have continued to decrease, it has become relatively time-consuming and expensive (Bau et al., 2009, Hodges et al., 2007). Technological advances have now made it possible to quickly and cost-effectively capture the identified target region, circumventing the more traditional approach of long-range PCR. These methods enable selective capture of the target sequence using either a custom-made hybridization array or more recently (released in February 2010), using an in-solution based method. Both methods enable targeted selection of sequences using capture probes: The array based method uses probes

which are attached to an array to capture the targeted sequence: The in-solution based method uses probes which are magnetically charged, so the target sequence can then be captured using magnetic beads.

When this sequencing experiment was undertaken, the only available sequence capture technology for custom targeted selection was the NimbleGen array (<http://www.nimblegen.com/seqcap/>). This technology has been used in numerous targeted re-sequencing projects which have led to the successful identification of disease-causing mutations (Rehman et al., 2010, D'Ascenzo et al., 2009). The array is capable of capturing up to 5 Mb of contiguous or discontinuous sequence using >60 bp probes, and requires a minimum input of 5µg of gDNA. Although this method offers an efficient and cost-effective alternative approach to PCR-based methods for targeted sequence selection, it is not without problems which can arise during the array design. The array is designed so that none of the probes are overlapping and also repeats are masked so that only unique regions in the genome are captured. Therefore, in contrast to traditional PCR-based methods which potentially provide 100% coverage of the target region, selective capture coverage of the target region, with either the array-based or in-solution sequence capture, can fall significantly below 100%, resulting in gaps requiring analysis using traditional sequencing methods.

Analysis of next-generation sequencing data

One of the most challenging aspects of next-generation sequencing is data analysis, which can be laborious and demanding on computational power. The Roche FLX system is the only system to provide the end-user with an integrated Windows compatible analysis suite, 'Genome Sequencer Analyzer'. This analysis suite is fully integrated with the sequencing platform, performing automatic trimming of reads and thus making the FLX system an attractive approach. The software has three applications; 'GS *De Novo* Assembler' which generates a consensus by *de novo* assembly of the shotgun sequencing reads, 'GS Reference Mapper' that maps the shotgun reads against a given consensus, producing a list of high-confidence mutations, and 'The Amplicon Variant Analyzer' for computing the alignment of

reads from amplicon libraries, identifying differences between the reads and reference sequence.

An analysis pipeline incorporating GS Reference Mapper software was best suited to the analysis of the sequencing data generated here, by mapping the shotgun reads to the reference consensus sequence (EquCab2) in order to identify a list of mutations. By comparing the reads to the reference, GS Reference Mapper generates a list of all high-confidence mutations using a combination of flow signal information, quality score information and stringent criteria to assign a mutation as a high-confidence difference. An example showing high confidence differences is shown in Table 5.2, using data from this sequencing experiment. For a variant to be deemed a high-confidence difference, it must be observed in at least three non-duplicate reads and have a minimum of five-fold coverage.

Start Position	End Position	Reference Nucleotide	Variant Nucleotide	Read Depth	Percentage of reads with observed variant
939	939	C	T	13	46%
940	940	A	G	13	100%
1523	1523	C	T	16	75%
1720	1720	C	T	9	56%
2634	2634	A	G	17	35%
2720	2720	T	C	15	40%
3185	3185	C	T	16	31%
4137	4137	C	T	3	100%
4233	4233	T	C	4	75%
4623	4623	A	G	11	45%
4670	4670	T	C	13	38%

Table 5.2: Example of an HCDiffs file exported from GS Reference Mapper software and opened in an Excel spreadsheet. For each high-confidence difference, the start and end position are given, along with the reference nucleotide and the observed variant. Also given is the read depth at the variant position and the percentage of reads in which the variant was observed.

With a list of mutations in-hand for each sample, the analysis pipeline then compares the genotypes of multiple samples to identify mutations which appear to segregate with the disease phenotype. Based on the number of times the variant is identified, which is given as a percentage of the reads which have the alternative variant

(preads), it can be inferred that the sample is either homozygous or heterozygous for the mutation. By inferring the genotype of a variant in multiple samples, a custom Perl script or query analyser can then be used to compare multiple samples for mutations which have the expected genotypic pattern. In theory, genomic variation in a diploid species is either homozygous or heterozygous and therefore the frequency of the observed variant is expected to be either 50% or 100% of the total reads. However, in practice the discrimination between homo- and heterozygous variants can be difficult, due to variation arising from sequencing errors, homopolymer length estimation errors, alignment errors, sampling variation, biological heterogeneity and low coverage (De Schrijver et al., 2010). Therefore, when inferring homo- or heterogeneity, a range of values should be used (e.g. heterozygote values ranging from 20 – 80%), to insure all potential variants are identified (I. Goodhead, University of Liverpool, unpublished observation). As a further check, to confirm true variants, the alignments of the reads to the reference sequence should also be visually inspected as this can be a good indicator of whether a variant is true or not.

Libraries for next-generation sequencing are prepared using fragmented gDNA, which creates difficulties when analysing the data for structural rearrangements, particularly when dealing with short single-end reads. When aligning the reads to the reference, reads which are shorter than the size of the rearrangement will all align to the same genomic position, thus giving increased read depth at the loci rather than indicating structural rearrangement. Furthermore, short reads may not span repeats, which can have a detrimental impact on mapping accuracy. The Roche Titanium Series has an average read length that exceeds any other platform, and is more likely to span short repeats, increasing the accuracy of mapping.

Detection of sequence rearrangements can be challenging with single-end reads, although by performing a *de novo* assembly the effectiveness of identifying mutations such as insertions and deletions is increased (Droege and Hill, 2008). However, the most successful detection of complex structural rearrangements and reliable repeat element mapping has been achieved using long paired-end reads (Fig 5.2) due to the two sequences in the pair having a known positional relationship to one another in the original genome (Fullwood et al., 2009). However, pair-end sequencing was not launched until April 2009 and so was not available for this

experiment (http://www.roche.com/media/media_releases/media_2009-04-23.htm).

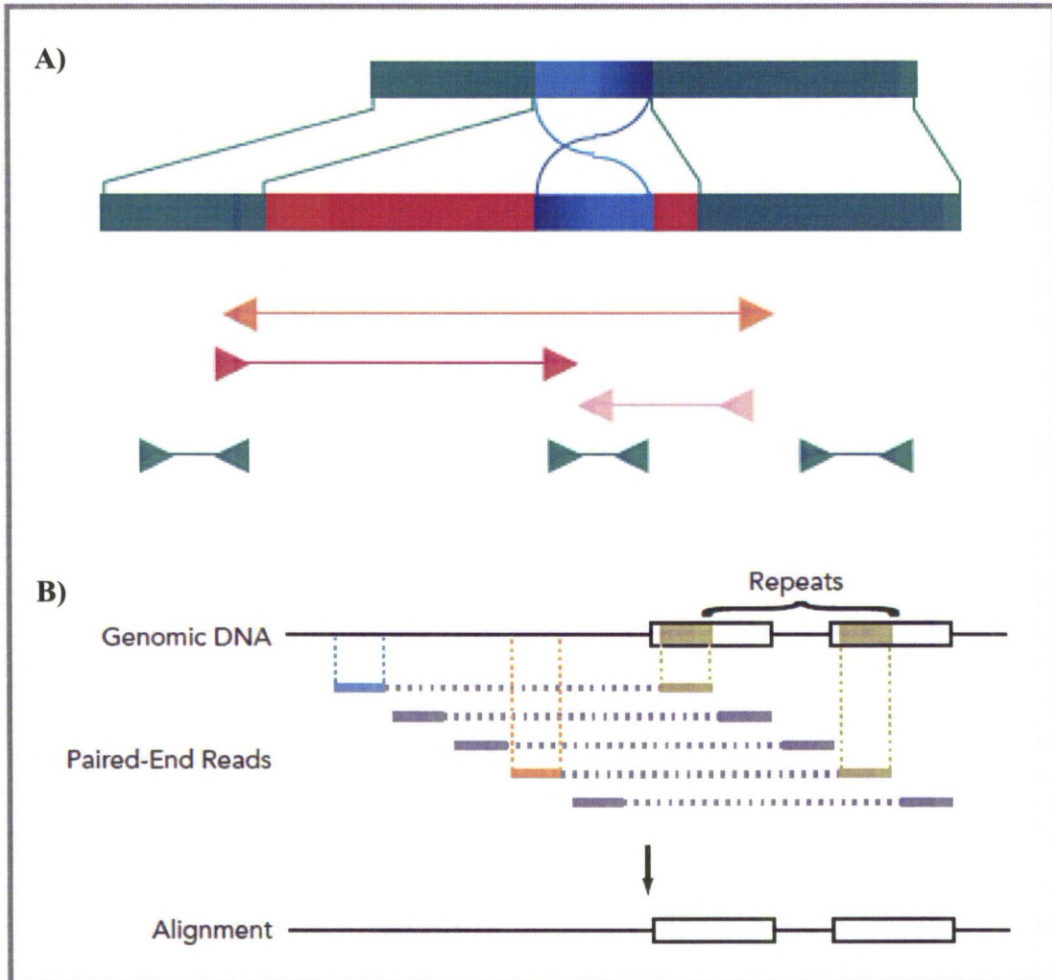


Figure 5.2: The benefits of paired-end reads: **A)** The complex rearrangement involves the inversion of the blue region, and the deletion of the flanking regions (red). Pairs of reads enable this to be detected: Colour coded arrow ends indicate the paired reads and the DNA fragments are shown by the coloured line. (Bentley et al., 2008) **B)** Reads in repeats (shown in green) can unambiguously be aligned to the genome as each read has a known positional relationship to its pair (blue and orange). Image reproduced from Illumina, USA (http://www.illumina.com/Documents/products/datasheets/datasheet_genomic_sequencing.pdf)

Objectives of re-sequencing the FIS critical interval to identify the causal mutation

Next-generation re-sequencing has proven to be very successful in mapping disease mutations in many species and therefore this approach was adopted to identify the FIS mutation. An array-based hybridisation method was used to capture the target region and the prepared libraries were then sequenced using the Roche 454 FLX Titanium series. The sequencing data was initially analysed to further refine the boundaries of the homozygous IBD FIS-affected haplotype and then analysed to identify sequence variants that segregated with the FIS phenotype. Potential candidate mutations were then interrogated by the testing of additional samples to identify the correlation of any disease causing mutation.

This chapter will describe:

- 1) The use of the re-sequencing data from an affected animal to confirm and further refine the boundaries of the homozygous IBD FIS-affected haplotype.
- 2) The identification of sequence variants that segregate with the disease, and the interrogation of these variants using additional samples to confirm the identity of the disease causing mutation.

5.2. Materials and Methods

Animals and Samples

Five Fell ponies (appendix 1) were selected for the re-sequencing project, based on sample performance in previous studies and observed haplotypes; these were one FIS-affected foal, three obligate carriers including the sire and dam of the FIS-affected foal, and a pony which was anticipated to be clear (non-carrier).

Details of the animals that were selected for next-generation sequencing project

The affected sample (03_08) was the animal with the smallest homozygous IBD haplotype, identified from the fine-mapping studies. The sire (15_08) and dam (14_08) of the affected animal were chosen as obligate carrier samples as they confirmed ambiguous genotype calls based on Mendelian inheritance. The third obligate carrier sample (13_08) was selected as it has a homozygous ‘affected’ haplotype. Sample 13_08 was confirmed as an obligate carrier, however previous genotyping (microsatellite mapping and SNP genotyping) revealed that this sample has the same haplotype as an affected sample. As it is known that this sample is an obligate carrier, it can therefore be presumed that this animal has the ancestral haplotype on which the mutation arose. The final animal which was selected was a predicted clear animal (25_01) which was identified during the PhD studies by Dr G. Thomas: genotyping data from this study supports his conclusion as it was homozygous for the wild-type allele throughout the critical interval.

DNA extraction from tissue and blood samples

Genomic DNA was isolated from tissue and blood lymphocytes using a revised version of the Nucleon® Blood Extraction kit (Appendix 2 and 4). DNA was re-suspended in ddH₂O.

Quality control procedures

Samples were assessed for DNA quantity and quality in accordance with the NimbleGen sample requirement and QC guide: Sample quantity was assessed using a Nanodrop, using a minimum threshold of 1.8 for the 260/280 absorbance ratio and 2.0 for the 260/230 absorbance ratio. The DNA samples were assessed for high molecular weight DNA by running 2µl of stock DNA on a 1% low-grade agarose gel at 100V for one hour. Samples were visualised under U-V light and assessed for evidence of low molecular weight (degraded) DNA. Samples with low molecular weight DNA were re-extracted.

Designing the sequence capture array

A custom tiling 385 k sequence capture array targeting a 3 Mb interval that encompasses the 992,356 bp FIS critical interval (Build EquCab2: ECA26: 28,942,655 – 31,942,655) was designed and manufactured by Roche NimbleGen (GenBank submission under study accession no. ERP000492). The array was designed using NimbleGen's standard repeat-masking algorithms for capture probe design. In order for the tiling array to be manufactured, NimbleGen were simply supplied with a FASTA file containing the raw sequence of the region of interest, which was downloaded from ENSEMBL.

Capturing the target sequence and sequencing the libraries

Sequence capture using the custom designed array and the subsequent sequencing was performed by the Centre of Genomic Research, University of Liverpool.

Overview of the sequence capture protocol

Initially a 3 - 5µg gDNA sample was fragmented using either sonication or nebulisation and quantified using a Bioanalyzer RNA 6000 Pico chip (Agilent Technologies, USA), to estimate the concentration of the library. The quantified DNA sample was then amplified by ligation-mediated PCR and assessed for quality using a Nanodrop spectrophotometer and for quantity on a Bioanalyzer DNA 7500

chip, prior to hybridisation. Hybridisation was performed at 42°C for ~64 hours before washing the microarray and eluting the captured gDNA samples. The eluted gDNA sample was then amplified by ligation-mediated PCR (LM-PCR) before assessing relative fold-enrichment by quantitative PCR (Fig 5.3). Finally, the library concentration was determined ready for emulsion-based clonal amplification and 454 sequencing.

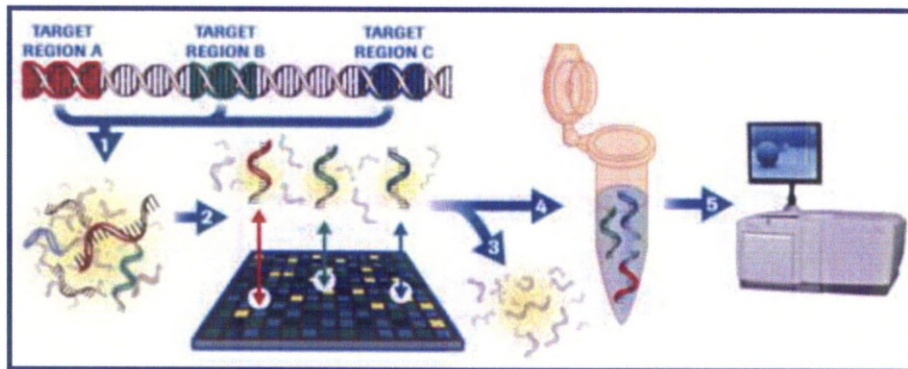


Figure. 5.3: An overview of the major steps for targeted sequence capture using the NimbleGen array based method, which takes approximately four days to complete: 1) The gDNA sample is fragmented by sonication or nebulisation. 2) The sample is hybridised to a NimbleGen Sequence Capture array. 3) Unbound fragments are washed away. 4) The target-enriched pool is eluted and LM-PCR amplified. 5) The enriched sample is ready for high-throughput sequencing, such as with a 454 Genome Sequencer FLX instrument. Image reproduced from Roche NimbleGen, USA (<http://www.nimblegen.com/products/seqcap/index.html>).

Overview of sequencing using the Roche 454 FLX Titanium series

The captured gDNA library was fragmented into small (300 to 800 bp) fragments and short adaptors were added to the fragment ends. The adaptors were complementary to the DNA capture beads, in order to ligate the single-stranded DNA library to the beads. Each individual bead carried a unique single-stranded DNA library fragment, ready for emulsion based PCR. The bead-bound library was emulsified with amplification reagents in a water-in-oil mixture and amplified within its own microreactor, excluding competing or contaminating sequences (Fig 5.4). Amplification of the entire fragment collection was then performed in parallel; resulting in several million clonal copies ($\sim 10^7$) of the fragment per bead.

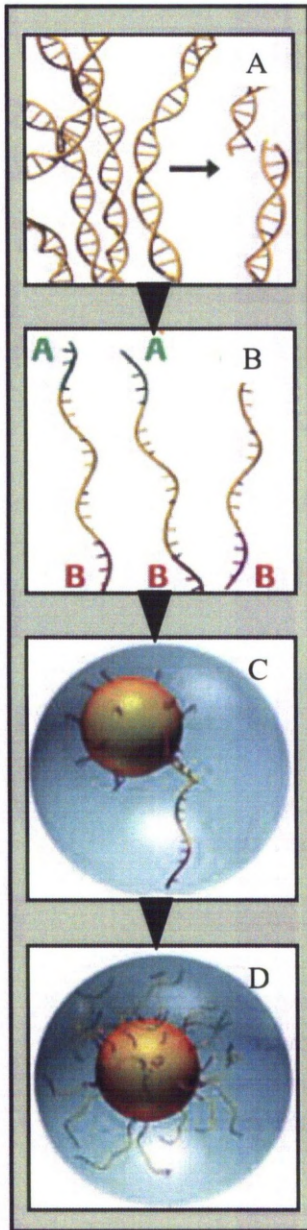


Figure 5.4: Preparation of the captured gDNA library for sequencing using the Roche 454 FLX Titanium series. A) The gDNA library is fragmented into small (300 to 800 bp) fragments. B) Addition of short adaptors to the ends of the gDNA fragments. C) Immobilization of the single-stranded DNA library to the beads. D) Emulsion-based clonal amplification results in several million clonal copies of the fragment. Images reproduced from Roche 454 (<http://www.454.com/enabling-technology/the-workflow.asp>).

When amplification was complete, the bead-bound library was loaded onto a picotiter plate for sequencing. The picotiter plate is designed so that each well can only house a single bead. After the addition of sequencing enzymes, nucleotides are flowed sequentially in a fixed order across the picotiter plate during a sequencing

run. During the nucleotide flow, hundreds of thousands of beads each carrying millions of copies of a unique single-stranded DNA molecule are sequenced in parallel. If a nucleotide complementary to the template strand is flowed into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal that is recorded by the instrument (Fig 5.5). The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow.

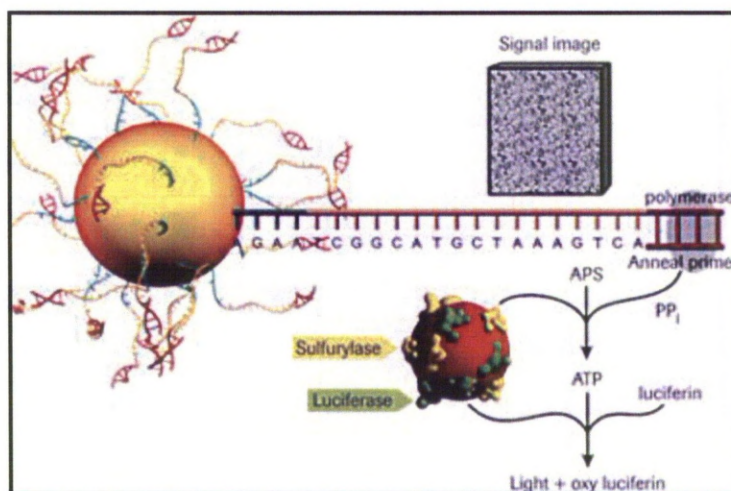


Figure 5.5: Sequencing reaction of the Genome Sequencer System. When a complementary nucleotide flows into the well, the polymerase extends the existing DNA strand by adding the nucleotide. Addition of each nucleotide results in a reaction which generates a light signal that is recorded and normalised to produce a flowgram for analysis. Image reproduced from Roche 454 (<http://www.454.com/enabling-technology/the-workflow.asp>).

Analysis of the sequencing data

Assembly of raw sequencing reads to the reference scaffold (Build EquCab2) was performed using GS Reference Mapper (Roche Newbler software) v. 2.0.00, using the graphical interface option. The automatic ‘NimbleGen SeqCap’ trimming parameter was selected and the GS read data sets (.sff files) imported into the assembly project and aligned to the reference sequence. A separate mapping project for each individual sample was created. The ‘454NewblerMetrics.txt’ output file for each alignment was used to assess sample performance using average read depth,

total number of reads and total number of aligned reads. Finally, the '454HCDiffs.txt' file was exported into an Excel sheet for further analysis. This file contained all of the 'High-Confidence' differences detected when the sequence was aligned to the reference scaffold. The high-confidence differences between the four samples with confirmed phenotype information were used for subsequent analysis.

Narrowing and confirming the critical region

Initially, the high-confidence difference file was examined in the FIS-affected animal in order to confirm the homozygous 992,356 bp critical region identified from the fine-mapping (chapter 4). All of the heterozygous variants (preads 20-80%) and homozygous variants with respect to the reference sequence (preads >80%) across the 3 Mb interval were identified in the high-confidence difference file from the affected animal. The repeat masking algorithm used by NimbleGen when designing the array was based on a human dataset so may not have masked all repeats, which undoubtedly led to some misalignments. In an attempt to reduce the number of ambiguous variants identified, additional repeat masking was performed, eliminating all those variants in the HC-Diff file that were known to fall within repeat regions. This was performed using a comprehensive horse-specific repeats file for the FIS target region which was generated by Dr. David Adelson at the University of Adelaide, Australia. Following this, all of the variants were verified using Sanger sequencing. Those which were confirmed as true variants were examined in additional affected animals and SNPs which were not verified were discarded from further examination. Verified SNPs were used to confirm that the boundaries of the 992,356 bp critical region were correct, refine the critical region, and also indicate the accuracy of the re-sequencing data.

Covering the gaps arising from probe design

Gaps in the sequence over the refined critical region which were due to lack of probe design were identified using SIGNALMAP version 1.7 (<http://www.nimblegen.com/products/software/>). These gaps were manually sequenced to maximise coverage. Initially the gaps were sequenced in an obligate carrier sample, and heterozygous mutations sought. Potential heterozygous variants were then be taken forward for verification together with variants identified by the 454 sequencing. Primers were designed to amplify the target sequence using Primer3

v4.0 online software as detailed in chapter 4 (page 97). Amplification of the target sequence was performed as detailed in chapter 4 (page 99) and the sequencing trace visualised using STADEN (<http://staden.sourceforge.net/>) to search for variants. For amplicon sizes >500 bp, the amplicon was also sequenced in the reverse direction, and where amplicon sizes exceeded 800bp internal sequencing primers were also used.

Investigating large scale rearrangements

To investigate large scale rearrangements, two-way visual alignments of each sample to the reference sequence were performed. Visual alignments were performed using the '454AllContigs.fna' files, which are automatically created by GS Reference Mapper software. Alignments were created using the web-based version (Abbott et al., 2007) (<http://www.webact.org/WebACT/home>) of the Artemis Comparison Tool (ACT) (Carver et al., 2005). The ACT alignment outputs were visually inspected for any ambiguous sequence differences between the experimental sample and the reference sequence.

To further investigate sequence rearrangement, sequence alignments for all of the samples were visually inspected in GS Reference Mapper, specifically examining for loss of coverage or gain of read depth in respect to the other samples.

Mining for candidate mutations

Candidate mutation mining was performed using MySQL query analyser (<http://www.mysql.com/>), using the '454HCDiffs.txt' files for the affected sample and the three obligate carrier samples. The '454HCDiffs.txt' file for each sample was imported into the database and labelled with the corresponding animal name (03_08, 13_08, 14_08 and 15_08). Homozygotes for the alternative allele were defined by preads >80% and heterozygotes were defined by preads 20-80%. Variants that were identified by the queries listed below were considered candidate mutations and compiled in an Excel spreadsheet for further interrogation. The following queries were used to mine for candidate mutations:

1) Variants which were homozygous in the affected animal (03_08) and heterozygous in all three obligate carriers:

```

selected* from affected_homozygous, 1308_heterozygous, 1508_heterozygous,
1408_heterozygous          WHERE affected_homozygous.Startpos =
1308_heterozygous.startpos AND affected_homozygous.Endpos =
1308_heterozygous.Endpos  AND affected_homozygous.AltSeq =
1308_heterozygous.AltSeq  AND affected_homozygous.Startpos =
1508_heterozygous.startpos AND affected_homozygous.Endpos =
1508_heterozygous.endpos  AND affected_homozygous.AltSeq =
1508_heterozygous.AltSeq  AND affected_homozygous.Startpos =
1408_heterozygous.startpos AND affected_homozygous.Endpos =
1408_heterozygous.endpos  AND affected_homozygous.AltSeq =
1408_heterozygous.AltSeq
    
```

2) Variants which were homozygous in the affected animal (03_08) and heterozygous in two obligate carriers. This query was performed to identify variants which would have been missed in the previous query if one of the carriers failed the '454HCDiffs' parameters (as stipulated by the GS Reference Mapper software) or was in fact a true heterozygote with a pread >80% or <20%. Therefore, this query was performed in the three possible combinations:

a) 03_08 compared to 14_08 and 15_08:

```

selected * from affected_homozygous, 1408_heterozygous, 1508_heterozygous
WHERE:
affected_homozygous.Startpos = 1408_heterozygous.startpos AND
affected_homozygous.Endpos = 1408_heterozygous.Endpos AND
affected_homozygous.AltSeq = 1408_heterozygous.AltSeq AND
affected_homozygous.Startpos = 1508_heterozygous.startpos AND
affected_homozygous.Endpos = 1508_heterozygous.endpos AND
affected_homozygous.AltSeq = 1508_heterozygous.AltSeq
    
```

b) 03_08 compared to 13_08 and 15_08:

selected* from affected_homozygous, 1308_heterozygous, 1508_heterozygous
WHERE:

affected_homozygous.Startpos	=	1308_heterozygous.startpos	AND
affected_homozygous.Endpos	=	1308_heterozygous.Endpos	AND
affected_homozygous.AltSeq	=	1308_heterozygous.AltSeq	AND
affected_homozygous.Startpos	=	1508_heterozygous.startpos	AND
affected_homozygous.Endpos	=	1508_heterozygous.endpos	AND
affected_homozygous.AltSeq	=	1508_heterozygous.AltSeq	

c) 03_08 compared to 13_08 and 14_08:

selected * from affected_homozygous, 1308_heterozygous, 1508_heterozygous
WHERE:

affected_homozygous.Startpos	=	1308_heterozygous.startpos	AND
affected_homozygous.Endpos	=	1308_heterozygous.Endpos	AND
affected_homozygous.AltSeq	=	1308_heterozygous.AltSeq	AND
affected_homozygous.Startpos	=	1408_heterozygous.startpos	AND
affected_homozygous.Endpos	=	1408_heterozygous.endpos	AND
affected_homozygous.AltSeq	=	1408_heterozygous.AltSeq	

3) Variants that were homozygous in the affected animal (03_08) and heterozygous in all three obligate carrier samples and within a protein-coding region. This query was performed to identify variants located within protein-coding sequence, and therefore of greatest interest:

selected* from affected_homozygous, 1308_heterozygous, 1408_heterozygous,
1508_heterozygous Exon_Boundaries WHERE affected_homozygous.Startpos =
1308_heterozygous.startpos AND affected_homozygous.Endpos =
1308_heterozygous.Endpos AND affected_homozygous.AltSeq =
1308_heterozygous.AltSeq AND affected_homozygous.Startpos =
1408_heterozygous.startpos AND affected_homozygous.Endpos =

```

1408_heterozygous.Endpos      AND      affected_homozygous.AltSeq    =
1408_heterozygous.AltSeq      AND      affected_homozygous.Startpos  =
1508_heterozygous.startpos    AND      affected_homozygous.Endpos   =
1508_heterozygous.Endpos      AND      affected_homozygous.AltSeq    =
1508_heterozygous.AltSeq AND (Affected_Homozygous.StartPos + 28942655) >
ExonStart AND (Affected_Homozygous.StartPos + 28942655) < ExonEnd

```

Interrogation of candidate mutations

Those variants that were identified from the various queries performed in MySQL were compiled as a list of candidate mutations for further interrogation. Initially, sequences containing the variant, with 300 bp of upstream (5') and 300 bp of downstream (3') flanking sequence were used to design primers to amplify the target sequence using Primer3 v4.0 online software as detailed in chapter 4 (page 97). Amplification of the target sequence (performed as detailed in chapter 4, page 99), containing the sequence variant was then performed in the four samples (03_08, 13_08, 14_08 and 15_08) to verify the sequence variant and that the correct genotypic pattern was observed (homozygous for the alternative allele in the affected animal and heterozygous in all three obligate carriers). Variants confirmed as having the correct phenotype-genotype pattern were then screened in a larger sample set of FIS-affected animals (38), obligate carriers (21) and animals with unknown phenotype (31) to interrogate the possible candidate mutations. To further interrogate any candidate mutations, they were also screened in 11 horse breeds, totaling 185 individuals (Thoroughbred, Appaloosa, Arab, Warmblood Sport Horse, Lipizzaner, Cleveland Bay, Dartmoor, Icelandic, New Forest, Sheltand and Shire), which were considered unlikely to have interbred with either the Fell or Dales and therefore should all be homozygous for the wild type allele.

5.3. Results

Sequence capture, sequencing and mapping the reads

Using a custom NimbleGen array with probes covering 92.91% of the 3Mb target region, DNA was enriched for five samples and the captured libraries sequenced using the Roche 454 FLX Titanium Series. A total of 2,618,719 reads were produced from the five samples, with an average length of 259 nucleotides, of which 1,786,802 (68.23%) aligned to the 3 Mb reference scaffold (EquCab2 assembly), after adaptor and quality trimming (Table 5.3). Sequence coverage was on average 34-fold, with sample 14_08 (obligate carrier) having the poorest performance with an average read depth of 13-fold.

The GS Reference Mapper software automatically detected high quality sequence variants (in accordance with the software algorithms) in the five samples (table 5.4). The high quality difference files were exported into Excel spreadsheets for further examination.

Sample ID	Total generated by 454 sequencing		Reads used for assembly after quality trimming		% used for assembly		Mean read depth of mapped reads	Mean length of mapped reads
	No. reads	No. bases	No. reads	No. bases	reads	bases		
25_01	294,447	82,855,909	238,500	68,859,899	81.00	83.11	23	289
03_08	596,735	149,916,608	387,458	111,553,865	64.93	74.41	37	288
13_08	624,844	158,160,278	400,074	116,677,291	64.03	73.77	39	292
14_08	187,430	50,559,839	132,855	38,929,151	70.88	77.00	13	293
15_08	915,263	238,158,874	627,915	178,738,114	68.60	75.05	60	285

Table 5.3: Summary statistics of 454 reads for the five samples that were used in the re-sequencing project.

Sample ID	Number of high-confidence differences identified
03 08	10,708
13 08	12,053
14 08	8,027
15 08	14,175
25 01	7,647

Table 5.4: Summary of the number of high-confidence differences identified by the GS Reference Mapper application.

Refining the homozygous critical haplotype

Initially high quality differences in the affected animal were used to confirm the identified 992,356 bp critical interval. Previously identified sequence variants which had been identified from SNP genotyping with the EquineSNP50 Beadchip were identified in the high quality difference file. There were 47 SNPs on the EquineSNP50 Beadchip which spanned the 3 Mb re-sequenced interval, and all 47 were verified in the next-generation sequencing data, thereby confirming the boundaries of the critical interval identified in chapter four.

After additional repeat masking, the 992,356 bp critical interval was examined for sequence variants that could potentially further refine the critical interval. A total of 44 heterozygous SNPs which spanned the critical interval (Appendix 7) were identified in the affected sample; 30 were ambiguous in the raw re-sequencing alignments while the remaining 14 further refined the critical interval to 842,542 nucleotides (29,928,903 Mb - 30,771,445 Mb). The 30 ambiguous SNPs were examined in the affected animal using Sanger sequencing. Of these, 22 failed verification; four failed due to amplification failure and 18 failed as they proved homozygous in the affected animal. The remaining eight SNPs were genotyped in 36 affected samples and further refined the critical interval: 26 samples were homozygous for all eight genotyped markers and the remaining 10 animals contained a smaller block of homozygosity, proximal and/or distal to but always containing a 375,043 nucleotide block of homozygosity (figure 5.6; ECA:30,372,577 Mb –

30,747,620 Mb). This homozygous interval was then thoroughly examined for the causal variant using the four samples with known phenotypes. The critical interval encompasses four genes: Intron 14-15 onwards of Intersectin-1 (*ITSN1*), ATP synthase subunit O (*ATP5O*), Sodium/myo-inositol cotransporter (*SLC5A3*) and mitochondrial ribosomal protein S6 (*MRPS6*).

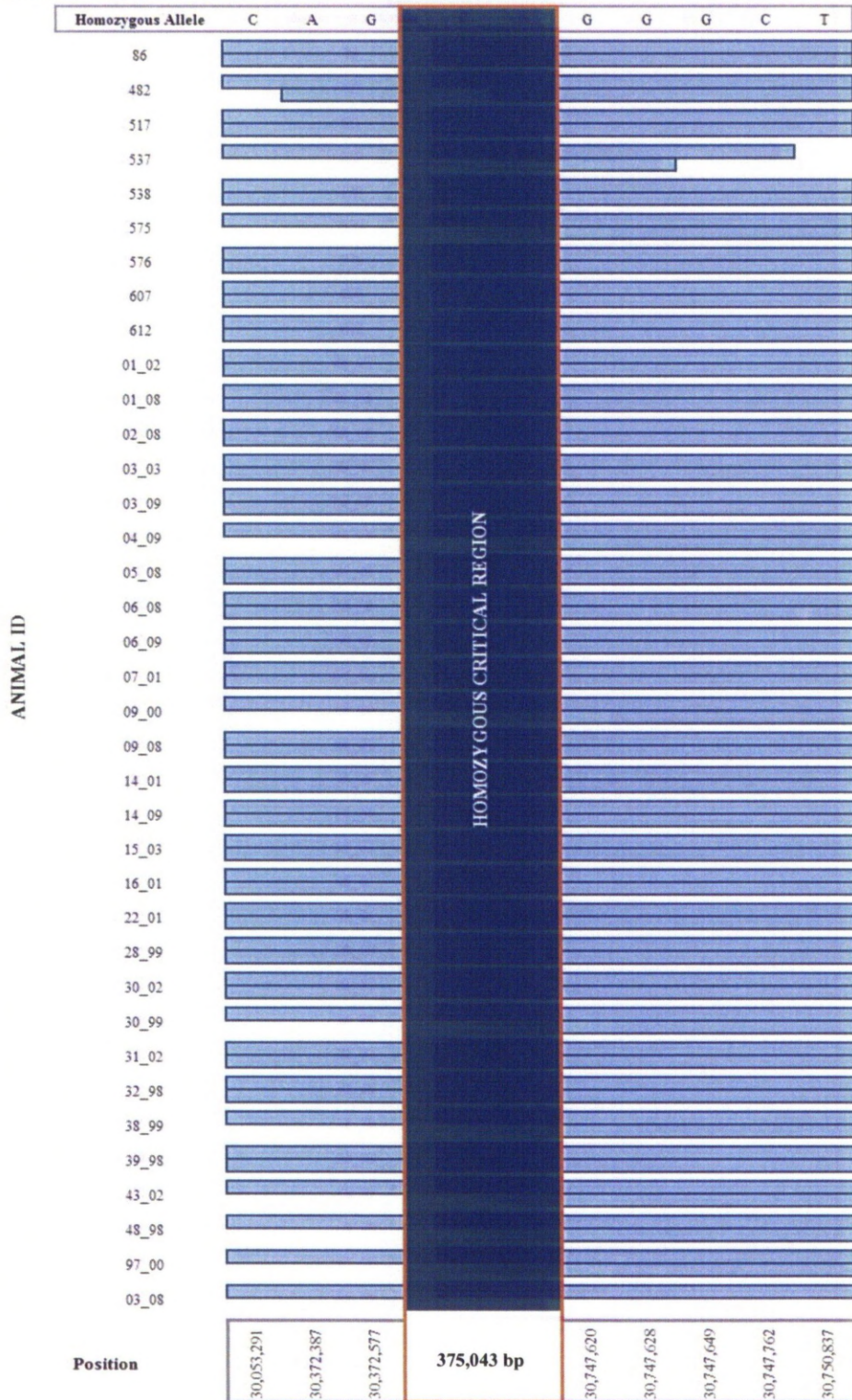


Figure 5.6: Narrowing the critical region. Haplotypes of the 37 affected animals which were used to further refine the homozygous haplotype for mutation screening. The smallest homozygous haplotype was a 375,043 bp interval in sample 03_08. Blue bars represent inheritance of the affected SNP as shown at the top of the diagram. The dark blue bar (outlined in orange) depicts the homozygous affected haplotype which was shared by all 37 affected animals.

Average read depth across the 375,043 bp critical interval was calculated for the four samples with known phenotypes. Read depth was calculated at 3,000 bp intervals, with the average read depth across all four samples being 36.21 (Fig 5.7). Read depth varied between the samples, with all four samples having a similar profile across the critical interval.

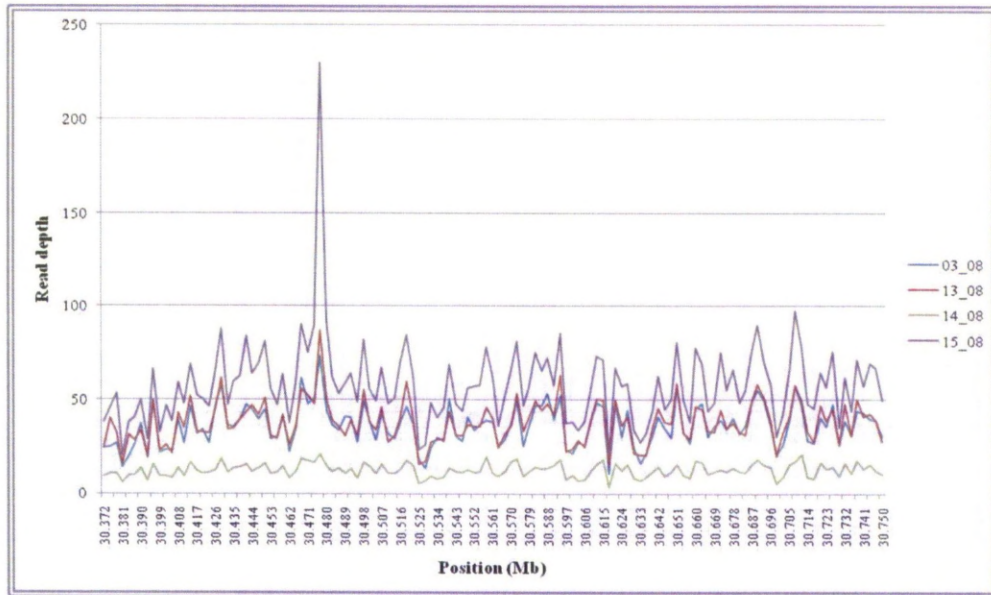


Figure. 5.7: Average read depth: The graph shows the average read depth for each sample across the 375,043 bp critical interval that was examined for the causal variant. Average read depth was calculated at 3000 bp intervals.

Manual sequencing to bridge those sequencing gaps that arose from the NimbleGen sequence capture probe design

Probe design for the NimbleGen sequence capture array provided 92.86% coverage of the 375,043 bp critical interval (Fig 5.8). Of the 7.14% of missing sequence, 3.05% fell within one of the four genes in the critical region, but none was in a protein-coding sequence and 4.09% was intergenic sequence. This missing sequence consists of 190 gaps of varying sizes in the critical region, averaging 190 bp (maximum 3370 bp and the smallest being 2 bp). In an attempt to increase coverage of the critical region, gaps in sequencing data were bridged with PCR then sequenced in animal 13_08. Sanger sequencing of these gaps increased coverage of the critical region to 98.36%. No heterozygous variants were identified during the sequencing of this obligate carrier. Of the 1.64% (27 gaps) which remained un-

sequenced, 1.04% (12 gaps) are within intronic sequence, with none falling within 200 bp of protein coding sequence and none falling within 15 kb of the first exon of any of the genes within the critical interval (Table 5.5).

Gene Position	Start	Finish	Gap Size	Position in relation to gene
ITSN1 (30,335,465 - 30,480,305)	30376387	30376425	38	Intron 15-16 (30374895 - 30377031)
	30380256	30380349	93	Intron 17-18 (30377410 - 30384321)
	30403820	30403832	12	Intron 23-24 (30403252 - 30411225)
	30432906	30432924	18	Intron 32-33 (30432707 - 30451172)
	30475028	30476083	1055	Intron 35-36 (30459170 - 30460961)
	30476302	30478171	1871	Intron 42-43 (30,473,303 30,480,156)
	30479102	30479591	489	Intron 42-43 (30,473,303 30,480,156)
ATP5O (30494968 - 30503454)	30495851	30496003	152	Intron 5-6 (30495468 - 30497793)
Intergenic	30512610	30512644	36	
	30550251	30550300	51	
	30592734	30592761	29	
	30599682	30599773	93	
	30604159	30604396	237	
	30608544	30608566	22	
	30617614	30617900	288	
	30626011	30626064	53	
	30626140	30626180	40	
	30631707	30631885	180	
30635027	30635716	689		
MRP36 (30686445 - 30704850)	30696448	30696498	50	Intron 1-2 (30686610 - 30704603)
	30697761	30697761	1	Intron 1-2 (30686610 - 30704603)
	30697979	30697996	17	Intron 1-2 (30686610 - 30704603)
	30699785	30699815	30	Intron 1-2 (30686610 - 30704603)
	30699895	30699936	41	Intron 1-2 (30686610 - 30704603)
Intergenic	30713380	30713420	42	
	30713741	30714217	476	
	30715459	30715506	47	

Table 5.5: Remaining gaps in the 375,043 bp critical interval. Of the 190 gaps arising from NimbleGen probe design, 27 remain. This leaves 1.64% (6150bp) unexamined. None of the gaps were within protein coding sequences but 1.04% were within intronic sequences (gaps within intronic sequences are highlighted in blue).

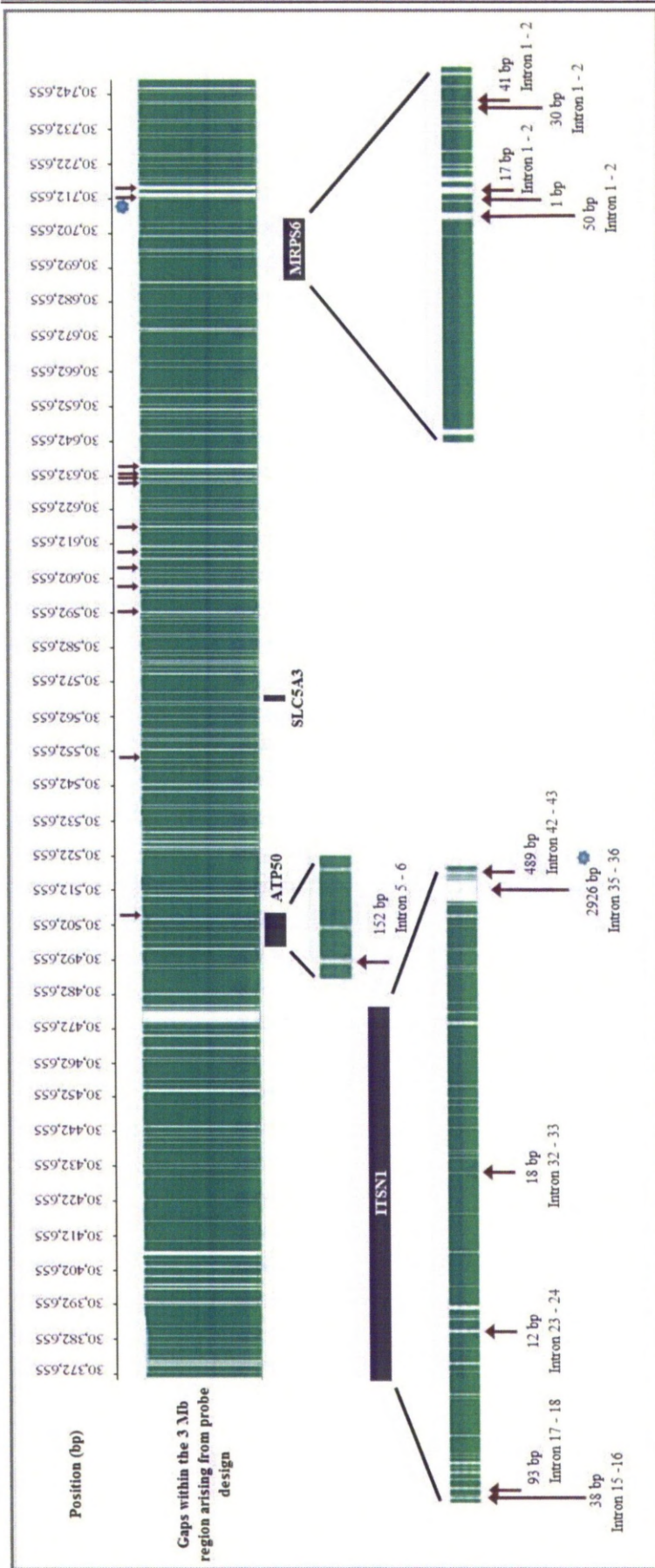


Figure 5.8: Image depicting the gaps within the sequencing data which arose from the probe design and their closure through manual sequencing. Purple blocks represent genes that are within the critical region and the sequencing gaps within these. All gaps arising from probe design were covered except those indicated by a red arrow: 100% coverage of all protein-coding sequence was achieved, with no intronic gaps exceeding 489 bp (with the exception of one gap within intron 35 - 36 IISN1). The sizes and introns in which these gaps lie are indicated. None of the intergenic gaps exceed 689 bp. Those arrows shown with a blue star represent a single gap that through manual sequencing was divided into two smaller gaps.

Investigation of large scale rearrangements

Alignments using Artemis Comparison Tool did not reveal any large scale rearrangements, with the affected sample showing an excellent sequence match to the reference sequence. Gaps in the sequence data were observed in all of the samples, when comparing them to the reference scaffold. Closer inspection revealed that these sequence gaps had arisen from probe design (Fig. 5.9).

Also observed were some sequence gaps (small deletions) at the contig boundaries, which were in all of the experimental sequence data, so not deemed as requiring follow-up. Gaps at the contig boundaries are expected with ACT sequence alignments, as they arise from the assembly process; small repeat regions cause the contig to collapse, as the repeat sequence stacks on top of one another at the contig boundary. The assembler is then unable to extend the contig further so it closes the contig and starts a new one, forming a small sequence gap between the contig boundaries (H. Browne, personal communication).

To further investigate sequence re-arrangements, in particular deletions including those identified by the ACT at the contig boundaries, the alignments were visually inspected in GS Reference Mapper. Manual comparison of the four samples with known phenotypes, did not reveal any significant sequence differences.

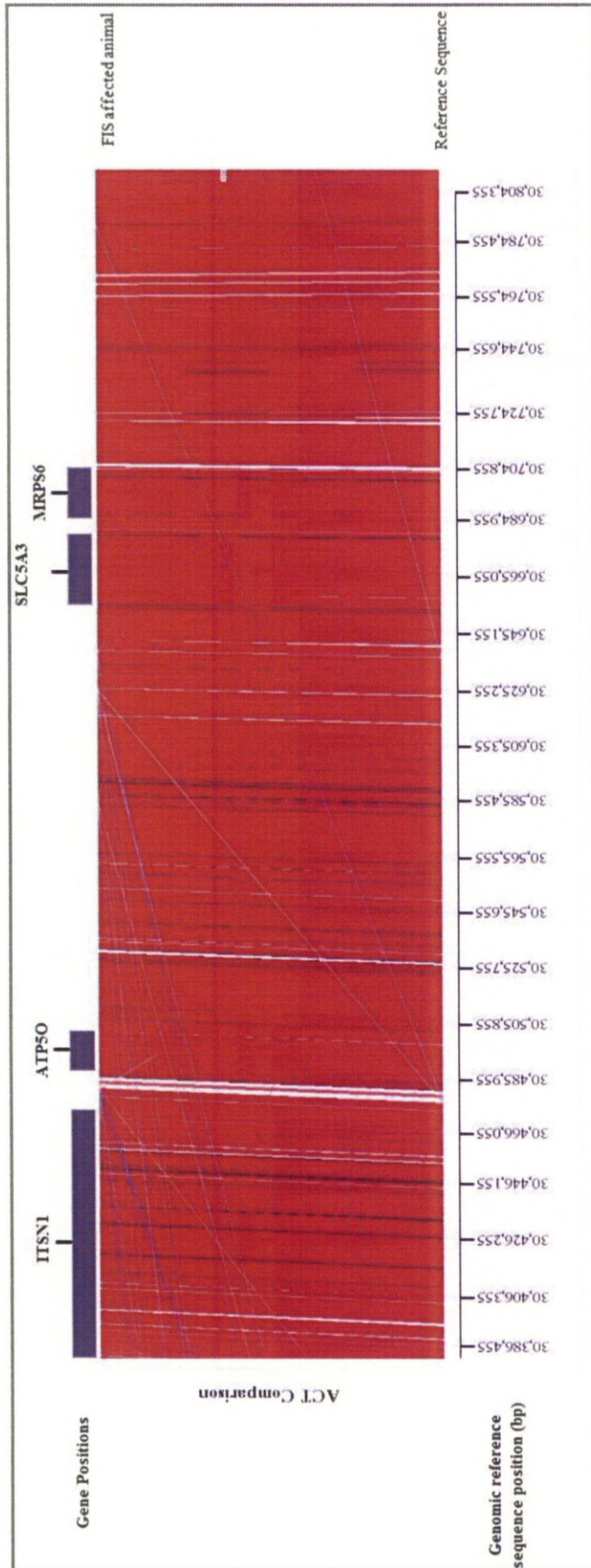


Figure 5.9: Comparison of the FIS-affected animal sequence to the reference genome sequence for the FIS critical region using Artemis Comparison Tool to look for sequence rearrangements. There was an excellent match between the reference sequence and the affected animal, providing no evidence of sequence rearrangements. Red blocks link similar regions of DNA with the intensity of the red colour being directly proportional to the level of similarity. Blue lines link regions inverted with respect to each other. White regions indicate missing sequence caused either by missing probe design or small tandem repeats which result in reads stacking on top of each other and causing a break in the contig. Black lines represent blocks of synteny used for sequence comparison and do not represent sequence variation.

Identification and interrogation of sequence variants in the critical region

Eleven variants, all of which were novel, were identified within the 375,043 bp critical interval from the data queries performed in MySQL (Table 5.6). Of these ten were SNPs and one was a three base-pair insertion were identified. Of the eleven variants, three were homozygous for the wild type-allele (highlighted in blue on table 5.6) in sample 25_01, which, based on previous genotyping information, was expected to be homozygous clear for the mutation.

Initially, primers were designed to amplify the target variant so that the genotypes could be confirmed in the sequenced samples. Only one variant, a SNP, survived further inspection and the remaining ten variants were excluded from further analysis: four SNPs were excluded as manual sequencing revealed that no variant existed at that locus, two SNPs were excluded as the correct genotype-phenotype pattern was not observed, the three base-pair insertion was excluded as manual sequencing revealed that there was no insertion at this locus and one further SNP was excluded as it was homozygous for the wild type allele in the affected sample and in an obligate carrier. The remaining two SNPs could not be excluded based on Sanger sequencing as the amplicons failed to sequence, however neither fell within protein-coding sequence (Table 5.6).

The SNP that survived further inspection and exclusion, was a novel SNP. The SNP (C>T) at 30,660,224 bp is within the protein coding sequence of *SLC5A3* and is predicted to cause a missense mutation (a nonsynonymous mutation is which a single nucleotide is changed, resulting in a codon that codes for an alternative amino acid) causing an amino acid substitution from proline to leucine.

Genomic Position		Variant		03_08 (Affected)		13_08 (Carrier)		14_08 (Carrier)		15_08 (Carrier)		25_01 (Unknown carrier status)		Is the variant within a gene:		Sanger Sequencing Results				
StartPos	EndPos	RefSeq	AltSeq	Nreads	Preads	Nreads	Preads	Nreads	Preads	Nreads	Preads	Nreads	Preads	Gene	Position within gene	03_08	13_08	14_08	15_08	
30,403,703	30,403,703	G	C	18	72	33	61	7	71	45	60	13	77	IITSN1	Intron 23-24	GG	GG	GG	GG	
30,407,557	30,407,557	C	T	29	76	52	63	-	-	72	65	19	89	IITSN1	Intron 23-24	-	-	-	-	
30,420,188	30,420,188	C	T	16	94	30	67	10	40	38	53	-	-	IITSN1	Intron 27-28	TT	TT	CT	CT	
30,524,595	30,524,595	G	C	11	73	19	63	-	-	26	54	7	43	-	-	GG	GG	GG	GG	
30,529,341	30,529,341	C	T	4	75	11	45	-	-	17	65	-	-	-	-	-	-	-	-	
30,529,406	30,529,406	G	A	8	87	15	67	-	-	22	68	5	80	-	-	GG	GG	GG	GG	
30,632,474	30,632,474	-	C:AG	6	83	9	44	10	40	16	50	-	-	-	-	No Insertion	No Insertion	No Insertion	No Insertion	
30,649,227	30,649,227	C	T	11	91	8	62	6	50	24	33	5	60	-	-	TT	TT	CT	CT	
30,660,224	30,660,224	C	T	32	91	36	39	16	44	51	37	-	-	SLCSA3	Exon 1	TT	CT	CT	CT	
30,696,518	30,696,518	G	A	4	75	13	54	-	-	23	52	9	33	MRPS6	Intron 1-2	GG	GG	GG	GG	
30,697,396	30,697,396	G	A	21	71	30	63	9	44	53	60	18	72	MRPS6	Intron 1-2	GG	GG	-	-	

Table 5.6: Candidate mutations that were identified from querying the high-confidence difference files in the five sequenced samples. The allele in the reference sequence is shown as 'RefSeq' and the variant shown as 'AltSeq'. For each sample, the number of reads (Nreads) at the position of the variant and the percentage of the reads that had the alternative allele (Preads) are given. Those shown in blue indicate samples where the variant was not observed in the high-confidence difference file as the variant failed the high-confidence difference parameters. The position of the variant is indicated if it falls within a gene. The results from the Sanger sequencing of the variants are shown with orange being indicative of a variant not being fully interrogated due to sequencing failure.

SLC5A3 mutation screening

To further interrogate the identified mutation, the SNP was screened in a larger sample set consisting of 21 obligate FIS-carriers, including the sire and dam of the affected Dales foal, 38 FIS-affected samples, including the Dales affected foal, and 31 Fell Pony samples with unknown carrier status. The SNP was homozygous for the alternative mutant allele in all 38 affected samples, it was heterozygous in all 21 obligate carrier samples and of the 31 Fell Pony samples tested with an unknown phenotype, two samples failed to work, 11 were heterozygous and 18 were homozygous for the wild type allele.

To further confirm this mutation as a plausible candidate, the SNP was screened in a selection of horse breeds (11 breeds; 184 individuals) which were considered unlikely to have interbred with either the Fell or Dales, and all proved homozygous wild-type (Table 5.7).

Breed	Number of samples screened	Sequencing result
Thoroughbred	29	Homozygous wild-type
Appaloosa	8	Homozygous wild-type
Arabian	21	Homozygous wild-type
Warmblood sport horse	17	Homozygous wild-type
Lipizzaner	2	Homozygous wild-type
Cleveland Bay	20	Homozygous wild-type
Dartmoor	19	Homozygous wild-type
Icelandic	8	Homozygous wild-type
New Forest	20	Homozygous wild-type
Shetland	20	Homozygous wild-type
Shire	20	Homozygous wild-type

Table 5.7: Screening of 184 individuals from 11 horse breeds, to assess the *SLC5A3* mutation in these populations. All 184 animals were homozygous wild-type (CC).

5.4. Discussion

The aim of this experiment was to confirm and narrow the FIS critical interval as much as possible and then comprehensively screen this interval to identify the disease causing mutation. The critical region was successfully defined as a 375,043 bp region which encompassed four genes; *ITSNI*, *ATP50*, *SLC5A3* and *MRPS6*. Interrogation of the critical interval led to the identification of a single variant which is highly associated with FIS, segregating 100% with the disease phenotype and being heterozygous for known carriers and homozygous for FIS foals.

Five animals were used for the re-sequencing experiment, four with confirmed phenotypes, and one with an unknown phenotype (25_01). Although this sample performed well, it could not be used to reject variants as the disease status of this animal could not be confirmed. Based on this, in hindsight, this animal should have not been re-sequenced as its use in the experiment was limited. The average read depth achieved across the entire 3 Mb interval for all five samples was ~34-fold; all of the samples performed well, providing high-quality sequencing data. The poorest performance was observed with sample 14_08, an obligate carrier that was the dam of the FIS-affected foal that was also sequenced. This sample achieved an average coverage of ~13-fold and of the three obligate carriers, this sample had the lowest number of high-quality differences (>4000 fewer). This low number of detected variants could be suggestive that low read depth limits the reliable detection of true variants because they fail to pass the strict variant calling threshold of GS Reference Mapper software.

For this experiment, a query analyser based analysis pipeline was used to identify possible candidate mutations, by comparing those variants identified in the HC-Diff file of samples with a known phenotype. HC-Diffs are only identified in regions of > 5X coverage, so as a consequence samples with poor read depth could result in the exclusion of a variant; it may be that the variant does exist, but due to poor read

depth it was filtered out. Potentially, this could hinder the detection of true variants when using a query based analysis pipeline, resulting in the loss of critical data. In an attempt to prevent the loss of potential candidates, multiple queries were performed, using the affected sample and different combinations of carrier samples. Only a single affected sample was sequenced so this individual was used in all of the comparisons, therefore it is possible that any variant in an area of poor coverage may have been missed. However this sample performed extremely well, achieving an average read depth of 37X, with closer inspection revealing that the only regions which fell below the minimum 5X threshold were those arising from probe design.

The reliability of SNP calling by the GS Reference Mapper software was examined by comparing the genotypes of those SNPs which were on the EquineSNP50 Beadchip to those calls made by the GS Reference Mapper. There were 47 SNPs on the Beadchip which spanned the 3 Mb critical region, and all genotypes agreed with the high-quality difference file. This not only provided evidence of the reliability of the software for identification of sequence variants but also enabled additional confirmation of the 992,356 bp critical interval described in chapter four. In an attempt to further narrow the FIS homozygous IBD haplotype, additional heterozygous variants were sought in the affected animal. As the repeat masking used for the NimbleGen probe design is based on a human algorithm, additional repeat masking was applied to limit the number of ambiguous SNPs that were identified for verification. Subsequently 44 SNPs were identified within the 992,356 bp region; visual inspection of the sequence alignment revealed that 30 were ambiguous so required further verification. Of these, 18 were not true variants, four were excluded as the amplicon failed to sequence, and the remaining eight were true heterozygotes. The eight heterozygote SNPs were sequenced in 37 animals which led to the critical region being further narrowed to 375,043 bp (30,372,577 – 30,747,620).

Coverage of the 375,043 bp critical interval, based on NimbleGen probe design, was 92.86%; 100% coverage is unachievable due to repeat masking. Repeat masking for probe design is essential to avoid non-specific hybridisation, which would result in the capture of repeated regions of the genome. To increase coverage of the target

region, Sanger sequencing was used to sequence these gaps. This increased coverage to 98.36%, with none of the remaining gaps falling within protein-coding sequence. The 1.64% of missing sequence comprises 1.04% intronic sequence, none of which falls within 200 bp of protein coding sequence. The remaining 0.60% is intergenic sequence, none of which falls within 15 Kb of the first exon of a gene in the critical interval. Although it is highly probable that the mutation causing a high-impact Mendelian disease like FIS is likely to fall within exonic sequence, it is now widely recognised that intergenic and intronic sequence can play a vital role in chromosomal structure and gene regulation through enhancers and regulatory regions. Therefore, although this experiment provided comprehensive cover of the critical region for mutation screening, this 1.64% of missing sequence cannot be conclusively excluded as harbouring the mutation.

The reliable detection of large scale rearrangements is notoriously difficult with single-end reads but at the time this experiment was conducted, paired-end reads were not an option. *De novo* assembly of single-end reads can be used to detect sequence rearrangements, but after personal communications with H. Browne at the Sanger centre, an alternative approach was adopted. This approach aligned the re-sequencing data to the reference sequence, using the ACT comparison feature, to visually inspect for possible rearrangements, duplications or insertion/deletions. There was an excellent match between the affected animal and the reference sequence and the alignments for the affected and obligate carriers were very similar, providing no evidence for significant rearrangement, duplication or insertion/deletion within the sequences. The only observed difference between the reference sequence and the re-sequencing data, in addition to sequencing gaps arising from probe design, was some small deletions at the contig boundaries. Deletions of this type are common with ACT comparisons and are due to the nature of the assembly. This is because small tandem repeat regions will tend to stack on top of each other, the assembler cannot extend the contig any further and the contig will be ended and another begun. To further screen for sequence rearrangements, the sequence alignment to the reference sequence was manually inspected in GS Reference Mapper. No significant differences were observed between the samples.

Therefore, the ACT comparisons and manual inspection provided no evidence to suggest that a large scale rearrangement was responsible for the FIS phenotype.

After maximum coverage of the target region had been achieved, and sequence rearrangements had been excluded as far as possible, the 375,043 bp critical interval was examined for potential candidate mutations. Interrogation of the sequencing data was performed using a query analyser, which examined and compared the high-quality difference files from those animals with a confirmed phenotype. The first comparison identified six sequence variants that were homozygous in the FIS-affected animal and heterozygous in all three obligate carrier samples. To prevent any potential candidates being overlooked due to poor sample performance, additional queries were performed to look for variants that were homozygous in the affected animal and heterozygous in just two of the obligate carrier samples. This query identified five additional variants for screening, all identified from the query which did not include sample 14_08 (the poorest performing sample).

The eleven variants were then sequenced using Sanger sequencing in the four re-sequenced animals, to identify if they were true variants with the expected genotypic pattern. This excluded eight of the variants as the expected genotypic pattern was not observed. Two further variants could not be excluded as candidates as the amplicon failed to sequence in all four animals; an intronic SNP in *ITSN1* and an intergenic SNP. The remaining variant, which was the only variant that fell within protein-coding sequence, had the expected genotypic pattern for a causal mutation. This variant was an exonic missense SNP in *SLC5A3* which causes an amino acid substitution from proline to leucine. This mutation segregates 100% with the FIS-phenotype; in addition other breeds which are highly unlikely to have interbred with either the Fell or Dales proved homozygous wild-type.

Although 100% coverage of the critical region was not obtained, the sequencing described here has successfully identified a missense mutation in *SLC5A3* which is highly associated with the FIS-phenotype. To confirm definitively that this is the disease causing mutation further studies would now be required to provide comprehensive coverage of the whole critical region and also to definitively confirm

that a sequence rearrangement had not occurred. Should these further studies reveal no additional candidate mutations as highly associated with the FIS-phenotype, functional studies would then be required to study the *SLC5A3* mutation. Nevertheless, the identified missense mutation segregates with the FIS phenotype 100% and is therefore deemed suitable as a genotypic test, to determine identify FIS-carriers. The FIS test was launched commercially in February 2010 and the testing results will be discussed in the following chapter. The possible transfer of this mutation into other breeds which are known to have interbred with the Fell and Dales Ponies over recent years will also be investigated.

Chapter 6

Population studies: Estimating the prevalence of FIS-carriers

	Page
Summary	150
6.1 Introduction	150
A population screen of the Fell and Dales breeds to estimate the prevalence of FIS-carriers	150
Population screen to assess the spread of FIS into other breeds	151
Pedigree analysis of Foal Immunodeficiency Syndrome affected foal	152
Aims and objectives of the FIS population studies	153
6.2 Materials and Methods	154
Population screen	154
Sample processing	156
Pedigree analysis of FIS-affected foals	159
6.3 Results	160
Population screen	160
Pedigree analysis of FIS-affected foals	161
6.4 Discussion	164

Summary

A single mutation which is highly associated with FIS has been identified and forms the basis of a DNA test which is currently being used to test for FIS-affected animals and FIS-carriers. This DNA-based test was used to perform a population screen of the UK Fell and Dales population, to estimate the prevalence of FIS-carriers and to also provide an estimate of the number of foals which are affected each year by FIS. Furthermore, this test was used to assess for the possible spread of the FIS mutation into other UK equine populations which were known to have interbred with either the Fell or Dales and so were considered 'at-risk'. This led to the identification of FIS-carriers in one other equine breed; further studies are now required to perform a large scale population screen of this breed to estimate carrier prevalence.

6.1 Introduction

A population screen of the Fell and Dales breeds to estimate the prevalence of FIS-carriers

To date, there have been no estimations of the prevalence of FIS-carriers amongst the UK Fell and Dales Pony populations. In 2001, it was suggested that the dramatic decline (15-25%) in Fell Pony foal registrations was mainly due to the loss of foals due to FIS (Bell et al., 2001). Based on Hardy-Weinberg Equilibrium (HWE), a 15-25% affection status would suggest that 47-50% of the adult Fell Pony population carry the FIS mutation. However, it is likely that this estimate of carrier prevalence is not accurate as the requirements for HWE are not met. For a population to remain in complete HWE, genotypic and allele frequencies must remain constant from generation to generation, which will only happen if the population is not subjected to specific population disturbances. Population disturbances which can disrupt HWE include; non-random mating, limited population size, selection, in-breeding and mutations (Hardy, 1908). The Fell and Dales Pony, which were founded with a relatively small population, are selected for mating based on breed standards and do

not breed randomly. Moreover, breeders (of Fell Ponies in particular) have become aware of certain breeding combinations producing syndrome foals in the past. As a result they have, in recent years, understandably, been avoiding certain matings which may result in the birth of an FIS-affected foal.

The identification of the *SLC5A3* mutation (as described in chapter five), has led to the development of an FIS-carrier test which is being used by breeders to screen their breeding stock. The data from the FIS testing laboratory will be presented in this chapter. This PCR-based test was also utilised to perform a population screen of the UK adult Fell and Dales Pony population to estimate the carrier prevalence. This requires that an unbiased (random) sample of each population is available.

The FIS testing laboratory data consists of samples submitted by owners and breeders for specifically for FIS mutation screening. However, the population screen will be performed using a separate cohort of samples which have been submitted for parentage profiling, and therefore will primarily consist of active breeding stock. Both of these datasets are liable to be biased, although estimating how biased is difficult, but they will provide a useful guide to FIS-carrier prevalence.

The FIS testing laboratory is also conducting the FIS test as a diagnostic test, to definitively confirm the diagnosis of foals which are suspected as FIS-affected. Use of this data will enable an estimate of the number of FIS foals born in 2010; this number is likely to dramatically decline over the coming years with informed breeding decisions now being possible. Both the Fell and Dales Societies have been contacted to provide data on the number of foals which have been registered over the past five years. Using this data, an average yearly foal crop can be calculated and the minimum disease prevalence calculated for 2010. An average number of foals (born in previous years) will be used for this calculation as foal registrations for 2010 will not be completed until early next year, which would be too late for inclusion in my thesis.

Population screen to assess the spread of FIS into other breeds

The spread of FIS into the Dales population was always a concern as both breeds have interbred over recent years and in June 2008, the first Dales FIS-affected foal was confirmed (Fox-Clipsham et al., 2009). However, interbreeding has not been

limited to these two breeds, as both have also interbred with several others over the last century. It was therefore deemed necessary to conduct a number of population screens to assess the spread of the FIS mutation into other ‘at-risk’ breeds. Discussions with breeders and various societies formed the basis for the selection of breeds for this screen. Additionally when performing pedigree analysis, all of these breeds were identified in early Fell and Dales pedigrees, confirming as best as possible that all of these breeds have interbred with the Fell and Dales and so are ‘at-risk’ from the spread of the FIS mutation. Breeds selected for this investigation included: Highland Ponies, Clydesdales, Welsh Section D (including part-bred), Exmoor Ponies and Coloured horses and ponies. In these discussions, all of these breeds were mentioned by multiple sources but one breed was repeatedly mentioned: the Coloured horse and pony. The Fell, Dales and traditional Coloured pony are very similar in type, all being sturdy and hardy ponies. Interbreeding between the Fell Pony and the Coloured pony has been very common over recent years, with several of the Fell Pony hill breeders regularly performing crosses. There was also anecdotal evidence of an FIS-affected Coloured foal in 2009 from a veterinarian who is very familiar with the presentation of FIS.

Pedigree analysis of Foal Immunodeficiency Syndrome affected foals

The FIS founder animal will be a common ancestor to all affected offspring, occurring on the maternal and paternal lineage of the affected animals. Pedigree analysis will enable the identification of the FIS common ancestor, which although likely, is not necessarily the founder of the FIS mutation. As FIS affects both Fell and Dales Ponies, it is likely that the common ancestor will have been actively breeding when these two breeds were interbreeding. Therefore, the common ancestor is likely to have been actively breeding between the 1940s, when the Fell Pony underwent its first genetic bottleneck and 1971, when the Dales inspection system was closed. Furthermore, the probable, though unproven, occurrence of FIS cases in the Fell Pony during the 1960s suggests that the common ancestor was breeding before this time. Here, pedigree analysis will be performed, tracing the lineage of all confirmed FIS-affected foals. By doing so, the common ancestor, and most probably the founder of the mutation, will be identified.

Aims and objectives of the FIS population studies

Population screens using sample archives submitted for DNA parentage profiling have successfully estimated the prevalence of inherited disease carriers amongst equine populations (Bernoco and Bailey, 1998, Swinburne et al., 1999). Based on this, a population screen using samples submitted to the Animal Health Trust for DNA parentage profiling will be used to estimate the prevalence of FIS-carriers amongst the UK Fell and Dales Pony population and also to assess the spread of this inherited disease into other equine breeds. Further to this, pedigree analysis will be performed, tracing the lineage of all FIS-affected foals, to identify the FIS common founder.

This chapter will describe:

1. The FIS-carrier prevalence in the UK Fell and Dales adult pony population, as determined from an anonymous and random screen of samples submitted to the Animal Health Trust for DNA parentage profiling.
2. The results obtained from the FIS testing laboratory, which performs FIS-carrier and affected foal testing. Using this data, an estimate of the number of foals born in 2010 that were affected by FIS will be calculated.
3. A population screen to assess the spread of the FIS mutation into five at-risk breeds which are known to have interbred with the Fell and Dales.
4. The identification of a single common founder stallion, which is present in the maternal and paternal lineage of all FIS-affected foals.

6.2 Materials and Methods

Population screen

All of the samples used in the population study were hair samples from adult ponies that were at least two years of age at the time of sampling.

Population study: Fell and Dales

Two hundred and fourteen Fell Pony samples and eighty-seven Dales Pony samples, which were among those sent to the Animal Health Trust for DNA parentage profiling between 2000 and 2010, were randomly and anonymously selected for FIS screening.

Population study of breeds which are known to have interbred with the Fell and Dales Ponies

Five breeds were selected as 'at risk', based on personal communications with breeds and discussions with the relevant breed societies. The five breeds selected were; Clydesdale, Exmoor, Highland, Welsh (section D and part-bred animals only) and Coloured horses and ponies (Table 6.1).

Breed	Number of samples	Sample source
Clydesdale	210	210 parentage profiling samples
Exmoor Pony	208	208 parentage profiling samples
Highland Pony	183	92 parentage profiling samples and 91 collected specifically for the purpose of this investigation
Welsh Section D (including Welsh part-bred animals)	210 (including 49 part-bred)	210 parentage profiling samples
Coloured Horses and Ponies	192	90 parentage profiling samples and 102 collected specifically for the purpose of this investigation

Table 6.1: *Summary of samples used for the population study. Breeds selected were those which are known to have interbred with the Fell and Dales Pony and so are at-risk from the spread of the FIS mutation into these populations. Parentage profiling samples were selected from those submitted to the Animal Health Trust. Additional samples were collected for the purpose of this investigation.*

Samples submitted to the Animal Health Trust for FIS screening

The FIS screening test was launched on the 1st February 2010 at the Animal Health Trust. Over the last 11 months (numbers correct as of 31st December 2010), a total of 888 samples have been submitted for testing, of which 186 were Dales ponies, 702 were Fell ponies and one is from a part-bred Dales pony.

All of these samples were processed by the Equine Parentage Laboratory at the Animal Health Trust using the protocol described below with the following amendments:

- In place of the MultiScreen PCR₉₆ 96-well filter plates for PCR purification, the QuickStep 2 PCR Purification Kit (EdgeBio, USA) was used in accordance with the manufacturer's instructions.
- In place of the isopropanol precipitation, DTR V3 96-Well Short Plates (EdgeBio, USA) were used for the removal of unincorporated dye terminators, in accordance with the manufacturer's instructions.

Sample processing

DNA Extraction

DNA was prepared from equine hair samples by placing six equine hair root bulbs into 0.2ml thin wall tubes (Alpha Laboratories, UK) in 90µl volumes of 0.5µl Tween 20 (Sigma Aldrich, UK), 0.6µl Tergitol NP-40 (Sigma Aldrich, UK), 10µl MgCl₂ (stock 25mM) (Promega, UK), 10µl 10 × DNA polymerase reaction buffer (Promega, UK), 2.7µl Proteinase K solution (Roche, UK) and 76.63µl ddH₂O. DNA was extracted into the buffer by heating on an MJ Tetrad PCR cycler (Bio-Rad Laboratories, Hercules, CA), at 60°C for 45 min, then 95°C for 15 min.

Polymerase Chain Reaction

DNA primers (forward primer with 18bp tail shown in bold 5'-**TGACCGGCAGCAA**AATTGCTCATGATTGTGGGGAGGATA-3'; reverse primer 5'-ATCAGGTTGGTCACATTCTGG- 3') which flank the *SLC5A3* mutation were used in a polymerase chain reaction (PCR) to amplify the 282 bp target region from the DNA sample. Amplification of the target sequence containing the informative SNP was performed as detailed on page 99 in chapter 4.

PCR purification

The PCR product was purified using MultiScreen PCR₉₆ 96-well filter plates (Millipore®, USA). PCR products were diluted in 200ul of ddH₂O and then transferred to the wells of the Multiscreen plate. The plate was placed on a vacuum manifold for 20 min. Following this, 20µl of ddH₂O was added to the individual wells and placed on a plate tilter for 10 min to re-suspend the product. Finally the

product was transferred to a 96-well skirted PCR plate (Axygen, USA) for storage at -20°C.

Sequencing reaction and precipitation of the products

The sequencing reaction was performed in 96-well PCR plates (Applied Biosystems, USA) in a 6µl volume, as detailed in chapter 4 on page 99, and loading onto an ABI3100 (Applied Biosystems, USA) for analysis according to the manufacturer's instructions.

Determining the genotype

Sequence traces were visualised using CHROMAS v1.45 (<http://www.technelysium.com.au/chromas.html>), to determine the base call of the nucleotide at position 134 (Fig 6.1). An animal which is homozygous for the wild-type allele 'C' is clear of the FIS mutation, whereas an affected animal is homozygous for the mutant 'T' allele and a carrier is heterozygous 'CT' (Fig 6.1)

CTCATGATTGTGGGGAGGATA TTTGTGGCTTTTATGGTGGTGATCAGCATTGCAT
 GGGTGCCAATCATCGTGGAGATGCAAGGAGGCCAGATGTACCTTTACATT CAGG
 AGGTAGCAGATTACCTGACGCCCC **C/T** GGTTGCGGCCCTGTTCTTCTGTCCATT
 TCTGGAAGCGCTGCAATGAACAAGGGGCTTTCTATGGTGAATGGCCGGCTT
 TTCTTGAGCAGTCCGTTTGACACTAGCCTTGCCTACCGTGCC **CCAGAATGTGA**
CCAACCTGAT

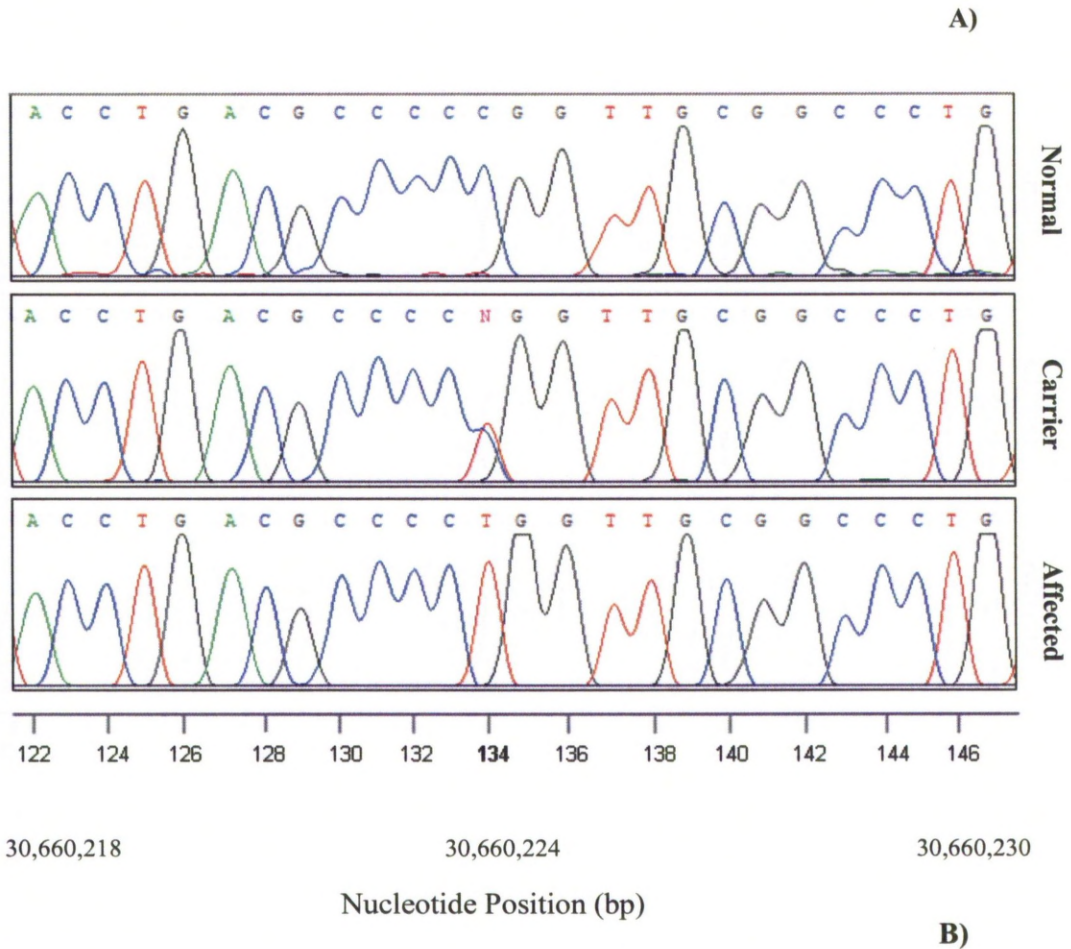


Figure 6.1: Single-base substitution which is highly associated with FIS. **A)** Target sequence which is amplified by the primers (shown in grey), to assess the genotype of the individual (SNP of interest is highlighted). **B)** Sequencing traces obtained from a normal (homozygous wild-type), FIS-carrier (heterozygous for mutant allele) and FIS-affected (homozygous for the mutant allele) pony. The mutant base is at nucleotide position 30,660,224 bp on chromosome 26 (position 134 in the amplicon).

Pedigree analysis of FIS-affected foals

Pedigree analysis, tracing the lineage of all confirmed FIS-affected Fell and Dales Pony foals, was performed using the online stud-books. Pedigree details for 38 confirmed FIS-affected foals were available: 16 (born between 1998 and 2003) had been collected as part of Dr. G Thomas's PhD studies, nine were collected by the author during this investigation (born between 2008 and 2009), and the remaining 13 foals were submitted to the FIS-screening laboratory for FIS diagnosis. Of these 38 samples, two were Dales Pony foals and the remaining 36 were Fell Pony foals.

6.3 Results

Population screen

Testing results from a population screen of the Fell and Dales

Using the newly devised DNA test for FIS, 214 Fell Pony samples were screened for the mutant allele. Of these, 38.32% (82 animals) were heterozygous for the mutant allele and therefore carriers of FIS. Of the 87 Dales Pony samples screened for FIS, 18.39% (16 animals) were carriers of the FIS mutant allele.

Testing results from a population study of five breeds which were considered at-risk due to the spread of the FIS mutation

Of the samples tested, there were no carriers identified in four of the five breeds (Clydesdale, Exmoor, Welsh and Highland) (Table 6.2). Of the 192 Coloured horse and pony samples tested, two (1.04%) were found to be carriers of FIS. Both of these samples had been submitted specifically for the purpose of this study. After discussions with the owners of these ponies, it was established that one was known to have Fell heritage and the heritage of the other pony was unknown.

Breed	No. samples tested	No. of carriers	% carriers
Clydesdale	210	0	0
Exmoor	208	0	0
Welsh (section D and part-bred)	210	0	0
Highland	183	0	0
Coloured	192	2	1.04

Table 6.2: A population screen to assess the spread of the FIS mutation into five breeds which are known to have interbred with the Fell and Dales.

FIS testing results for Fell and Dales samples submitted to the AHT for FIS screening

Since the FIS test was launched in February 2010 (11 months ago), 888 samples have been submitted for FIS screening, both for the identification of carrier animals and as a diagnostic tool for the diagnosis of suspected FIS-affected foals (Table 6.3).

The single part breed Dales pony which was submitted for FIS testing was homozygous clear. Of the 186 Dales pony samples tested, one was confirmed as an affected foal and 20 of the adult samples were identified as FIS-carriers. Of the 702 Fell Pony samples submitted for testing, 116 were Fell Pony foals and the remaining 586 were adult Fell ponies. Of these, 13 were confirmed as FIS-affected and 282 of the adult samples confirmed as FIS-carriers.

	Clear	Carrier	Affected	Total
Dales Pony testing data				
Foals	0 (0%)	0 (0%)	1 (100%)	1
Adults	165 (89.19%)	20 (10.81%)	0 (0%)	185
Fell Pony testing data				
Foals	54 (46.55%)	49 (42.24%)	13 (11.21%)	116
Adults	304 (51.88%)	282 (48.12%)	0 (0%)	586

Table 6.3: *Statistics from the FIS screening laboratory.*

Pedigree analysis of FIS-affected foals

Pedigree analysis identified 13 prominent sires which appeared in the maternal and paternal lineage of more than 6 of the affected foals and were used for breeding between 1940 and 1971. One sire which was born in 1939 (sire 13) was identified in the paternal and maternal lineage of all FIS-affected Fell Pony foals. A further sire,

born in 1946, was identified in the maternal and paternal lineage of all 38 foals, including the two FIS-affected Dales Pony foals (table 1.2). Pedigree analysis suggests that this sire was extremely popular for breeding, appearing multiple times in any one pedigree. In one pedigree this sire appeared eight times in an eight generation pedigree, being the great-grandfather, grandfather and father of a single offspring. Furthermore, he is also sire to three of the remaining 12 prominent sires; sires two, five and nine. From this it can be concluded that sire one, which was born in 1946 and was actively breeding in the 1950s and 1960s, is the common ancestor to all FIS-affected foals examined and is therefore likely to be the founder in whom this mutation arose.

Sire	FIS affected Foal		1998		1999		2000		2001		2002		2003		2008		2009		2010		2010		2010		2010		2010				
	Breed	Year of birth	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell	Fell		
1	1946	1946	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
2	1964	1964																													
3	1961	1961	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
4	1954	1954	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
5	1953	1953																													
6	1959	1959	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
7	1953	1953	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
8	1968	1968																													
9	1957	1957																													
10	1958	1958																													
11	1951	1951																													
12	1965	1965	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
13	1939	1939	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	

Table 6.4: Pedigree analysis to identify prominent sires which were actively breeding between 1940 and 1971. Thirteen sires were identified, of which one (sire one which is highlighted in blue) appeared in the maternal and paternal lineage of all 38 confirmed FIS-affected foals, of which two were Dales Pony foals.

6.4 Discussion

The aim of this investigation was to estimate the prevalence of FIS-carriers amongst the UK Fell and Dales pony population and to assess the spread of the FIS mutation into other at-risk breeds. This investigation has successfully estimated the prevalence of carriers amongst the UK Fell and Dales Pony population and has also provided data on other breeds, confirming that FIS-carriers exist in the UK coloured pony population.

Here, we conducted an anonymous and random FIS-carrier screen of samples which had been submitted to the Animal Health Trust for parentage verification. This data suggested that approximately 38% of the UK adult Fell Pony population and 18% of the UK Dales adult Dales Pony population were carriers of FIS. As this was an anonymous screen, there was no sex information known about those samples used in the study. However, it is known that both the Dales and Fell Pony Societies require all stallions to be registered, which involves submitting a DNA sample to the Animal Health Trust. Since January 2008, approximately twice as many stallion samples have been submitted for DNA-profiling as mare samples. It is highly likely therefore that this population screen is biased towards breeding stallions rather than giving a true reflection of the population as a whole.

In February 2010 the FIS-carrier test became publicly available. This test can be used as a carrier test or as a diagnostic tool for suspected FIS-affected foals, and has been in great demand for both uses over the past eleven months. Since the launch of the test, 771 adult samples have been submitted for carrier screening, of which 586 were from Fell Ponies and 185 from Dales Ponies. This data suggests that approximately 48% of the UK adult Fell Pony population and 10% of the UK adult Dales Pony population are carriers of FIS. Compared to the anonymous screen, this suggests that half as many Dales Ponies are carriers of FIS and that an additional 10% of Fell Ponies carry FIS.

It is likely that the testing results are also biased, and the carrier prevalence either inflated or deflated. Samples that are submitted for testing are selected by the owner; they may submit samples from animals which they suspect may have had an FIS foal in the past and want confirmation that the animal is a carrier, or alternatively they may submit samples only from those animals they hope to be clear as they have never had an FIS foal. In contrast to the Fell and Dales population screen, which only included UK adult samples, the FIS testing includes data on Fell and Dales ponies from outside the UK, including samples from across Europe, the United States of America and Canada. This data suggests that the FIS-carrier prevalence is higher and is statistically significantly different in the Fell pony than that estimated from the anonymous population screen, increasing from approximately 40 to 50%. Conversely, with the Dales pony it suggests a decrease in carrier prevalence, although not statistically significant, compared to the anonymous population screen, decreasing from approximately 20 to 10%. Again, it is likely that the testing results are biased towards breeding stock; however, it is possible that it provides a more accurate representation of the wider population as breeders will test both breeding mares and potential stallions, as well as stallions to be licensed. Furthermore, as testing numbers increase any bias will decrease, providing a more accurate indicator of carrier prevalence.

The FIS test has also been utilised as a diagnostic tool, to confirm suspected FIS-affected foals. Since the launch of the test, 116 Fell pony foals and one Dales foal have been submitted for testing. The single Dales foal was suspected of being FIS-affected, and this was confirmed using the DNA test. Of the 116 Fell pony foals tested, 13 (11.21%) were suspected as FIS-affected and were confirmed as such using the genotyping test. All of these foals were euthanized. Of the remaining 99 foal samples, 52 (46.43%) were FIS clear and 47 (41.96%) carriers. Of these, four were believed to be FIS-affected, and were submitted from the same breeder. Of the four animals, two were confirmed as FIS-carriers and the other two were clear of FIS. Conversations with the veterinary surgeon revealed that two of the foals survived, one later confirmed as having tetanus and the other chronic diarrhoea which subsided with treatment. Of the two foals which died, one had a post-mortem and had pathological findings consistent with *Rhodococcus equi* infection, and the

remaining foal died prior to being seen by a veterinarian and did not undergo a post-mortem.

Communications with both the Fell and Dales Pony Societies has revealed that both societies have seen a gradual increase in the number of foal registrations, over recent years. The Dales society has reported a gradual increase in foal registrations over the last five years; prior to this they had observed a gradual decline in registrations over the previous 10 years. It is thought that this increase coincides with the Dales pony being listed as 'critical' by the Rare Breeds Survival Trust, which encouraged additional breeding. The Fell pony society has observed an increase in foal registrations since the 1980s, with the first decrease in foal registrations being seen in 2010.

Over the past five years, Fell pony foal registrations have averaged at 487 foals per year and Dales foal registrations at 170 foals per year. Although it is very likely that not all FIS-affected foals were submitted to the Animal Health Trust for testing, we can conclude that in 2010 at least 2.7% (13/487) of Fell pony foals and 0.59% (1/170) of Dales pony foals born were affected by FIS. It is highly probable that this is an underestimate of the real affected statistic, as not all affected samples are likely to have been submitted for testing. Therefore, this data provides an estimate of the probable minimum FIS affection statistic for 2010. This affection statistic is substantially lower than that estimated 10 years previously by Bell *et al* (2001), who suggested that the dramatic decline in Fell pony foal registrations was mainly due to FIS and estimated that approximately 15-25% of foals born were affected by this disease. The results from this investigation suggest substantially fewer Fell Pony foals are now affected by FIS. It is unlikely that the number of foals born with FIS reduced this dramatically without the aid of a carrier test. It is more likely that the estimate calculated by Bell *et al* (2001) was elevated and that the dramatic decline in Fell foal registrations at that time was not only due to FIS but also due to the diminishing popularity of the Fell Pony due to the negative publicity arising from FIS.

In addition to performing a population screen of the Fell and Dales, a population screen was also conducted to evaluate the spread of FIS into other breeds known to have interbred with either the Fell or Dales in recent years. After discussions with

individual breeders and societies, five breeds were selected: Exmoor, Highland, Clydesdale, Welsh Section D (including part-bred) and Coloured horses and ponies. FIS-carriers were only identified amongst the Coloured samples, with two carriers out of the 192 animals tested. Both of the carriers were from samples collected specifically for the purpose of this study: One was known to have Fell Pony ancestry and the other was collected randomly at a horse show from an animal with no known Fell or Dales heritage. Of all the breeds tested, FIS-carriers were considered most likely amongst the Coloured population, due to the known Fell and Dales influence on this pony type. Further studies of the coloured population, using a larger sample set, would provide a more accurate estimation of the carrier prevalence. To prevent the further spread of this defect into the population and an increasing carrier rate, which would possibly lead to FIS-affected foals, breeders should be cautious in their breeding decisions. Owners of Coloured horses and ponies with known Fell or Dales heritage should consider screening their stock for carriers. Because the carrier prevalence in the Coloured population is still low, carriers of FIS should be excluded from breeding programs. Particular caution should be taken when cross-breeding with Fell or Dales ponies, using only those animals which have been proven clear of the FIS defect.

The data presented here suggests that approximately 38 – 48% of adult Fell ponies and 9 – 18% of adult Dales ponies are carriers of FIS. We can also say that at least 2.7% of Fell and 0.59% of Dales Pony foals born in 2010 were affected by FIS. These numbers are very high and it is therefore extremely timely that the carrier test is now available. The value of the test will be seen in the next few years as owners and breeders become selective in their choice of mating partners for their ponies and the carrier rate will eventually drop and the numbers of FIS foals should fall dramatically. I am therefore hopeful that in years to come, this dreadful mutation will be eradicated. This investigation has also confirmed that the Coloured horse and pony population is affected by FIS as carriers amongst the population have been identified. It is also worthwhile testing samples from other breeds as we have learnt from the histories of the Fell and Dales ponies that waiting to identify a sick foal means that a high carrier rate already exists in a given horse population. We must remain vigilant.

Pedigree analysis revealed a single common sire in the maternal and paternal lineage of all FIS-affected foals. This sire was actively breeding in the 1950's and 1960's, a time when the Fell Pony had recently undergone a genetic bottleneck and the Dales Society allowed interbreeding with the Fell Pony. It is highly likely that this sire is the founder of the FIS-mutation and his popularity enabled the rapid spread of this mutation into the population. Due to the relatively small breeding population and the influence of this sire, it was only a matter of time before a consanguineous mating occurred, giving rise to an FIS-affected foal.

Chapter 7

A pilot study to investigate transcriptional changes in FIS-affected foals

	Page
Summary	169
7.1 Introduction	169
Investigating global transcriptional changes	169
Why perform a pilot study?	172
The experimental design: A pilot study to investigate global transcriptional changes in FIS-affected foals	173
7.2 Materials and Methods	176
Animals and sample collection	176
Total RNA extraction from bone-marrow samples	177
RNA-Sequencing using the Illumina Genome Analyser II System	178
Analysis of RNA-Seq data: Detecting differentially expressed genes	180
Analysis of molecular interactions	183
7.3 Results	185
Animals and Samples: Selection of samples for RNA-Seq	185
Mapping the reads	186
Identifying differentially expressed genes	187
Pathway analysis using Ingenuity	202
7.4 Discussion	209

Summary

RNA-Seq is a revolutionary new technology which has been developed to provide a global, quantitative, survey of transcription in biological samples. This technology is highly accurate, requiring no previous knowledge of the genome and is capable of detecting both known and novel transcripts. RNA-Seq was therefore selected as the technology of choice for investigating global transcriptional changes in FIS-affected foals. Here, a pilot study was conducted to investigate the variables involved in performing RNA-Seq on bone marrow samples from FIS-affected foals and controls, and in comparing gene expression levels in both cohorts. This study has provided vital information regarding the best strategy for the design of future studies, and has yielded some highly valuable initial data. Functional groups of genes related to cellular growth, proliferation and development, the development of the haematological system, haematopoiesis and tissue development were identified as those pathways which were most significantly disrupted in FIS-affected foals.

7.1 Introduction

Investigating global transcriptional changes

Gene expression levels vary greatly during development, between cell and tissue types, and between healthy and diseased tissues. Therefore, investigating differential gene expression has become popular for answering many biological questions, especially for investigating differentially expressed genes between diseased and normal (Marioni et al., 2008). Historically, hybridisation based DNA microarrays have been used for investigating global gene-expression (Bright et al., 2009), but over the last few years, alternative high-throughput sequencing-based approaches have become available (Morrissy et al., 2010).

Although hybridisation-based methods are well suited to high-throughput and are relatively inexpensive, they have many limitations. One such limitation is that they

are subject to high levels of cross hybridisation, which in turn leads to high-background levels. This greatly limits the accuracy of expression measurements, particularly for those transcripts that are present in low abundance or high abundance (Wang et al., 2009). A further potential problem is that the probes can differ significantly in their hybridisation properties and therefore comparing results from different laboratories is hugely problematic, requiring complicated normalisation algorithms (Rosenkranz et al., 2008). Furthermore, the arrays are limited to interrogating transcripts for which probes have been included on the array, so does not provide comprehensive transcriptional analysis and cannot detect novel transcripts.

TAG-based sequencing methods utilise ‘short sequences’ or ‘sequencing tags’ to identify transcripts for expression profiling (Harbers and Carninci, 2005). TAG-based methods are a popular alternative to microarrays, providing precise digital gene expression levels though at a much higher cost. TAG-based methods include serial analysis of gene expression (SAGE) (Velculescu et al., 1995), cap analysis of gene expression (CAGE) (Shiraki et al., 2003) and massively parallel signature sequencing (MPSS) (Brenner et al., 2000). Although these methods differ, they are all based on the principle of using small fragments that correspond to sequences at either the 3’ or 5’ end of the transcripts, to provide a digital count of the number of times the transcript is present. However, these methods provide limited information as a significant proportion of the short tags cannot be mapped uniquely to the reference genome; they only provide information on a very small portion of the transcript and not all genes will necessarily contain the TAG sequence and are therefore not quantifiable (Wang et al., 2009).

The development of next-generation sequencing platforms has given rise to a new method for both mapping and quantifying transcriptomes, a method known as RNA sequencing (RNA-Seq). RNA-Seq enables the characterisation of gene expression in tissues by utilising high-throughput sequencing technologies to sequence transcribed complementary DNA (cDNA). This in turn enables comparison of diseased and normal tissue, to determine how transcriptional activity of the tissue is affected by the disease.

RNA-Seq involves the direct sequencing of cDNAs using a high-throughput next-generation sequencing platform, followed by mapping the sequencing reads to the reference genome (Nagalakshmi et al., 2010). This method provides an accurate and quantitative measure of individual gene expression and has very low, background signal because the majority of DNA sequences are mapped unambiguously to unique regions of the genome. One of the major limitations of microarray based experiments is accuracy as they lack sensitivity for genes which are expressed at very low or very high levels. In contrast, RNA-Seq experiments have been shown to be highly accurate and replicable, with little variation between technical replicates (Marioni et al., 2008). RNA-Seq also has the added advantage that it performs a global survey of the transcriptome, detecting novel transcribed regions in an unbiased manner as it requires no prior knowledge of transcribed regions (Wilhelm and Landry, 2009). Although RNA-Seq overcomes the limitations of both microarray and TAG-based methods, it too has limitations. RNA-seq has limited use with some experiments as the analysis requires an annotated genome sequence, and further to this, the technology is expensive and requires computational experts.

Consequently, due to its high accuracy and ability to detect novel transcripts, RNA-Seq was chosen as the technology of choice for investigating global transcriptional changes in FIS-affected animals when compared to healthy animals.

FIS and anticipated transcriptional changes

FIS is a lethal disease which predominant features include a severe B-lymphocyte deficiency and a profound anaemia. Based on this we would expect to see significant disturbances to those pathways which are involved in the development of B-lymphocytes and red blood cells. Haematopoiesis, which takes place in the bone marrow, is responsible for the formation of all cellular blood components, with all cellular components being derived from the same cell, the [haematopoietic](#) stem cell. Based on this, we would expect to see disturbance to those genes which are involved in the haematopoietic cell lineage. However of all the cells derived from the haematopoietic cell lineage, only B-lymphocytes and erythrocytes are reported to be affected by this disease and therefore we will be examining for gene disturbances to genes which are exclusively involved in the development of these two cells types.

From examining this pathway it is acknowledged that there are no genes which are exclusively involved in only the development of B-lymphocytes and erythrocytes, and therefore it is unlikely that the primary mutation has a direct impact on a specific gene which results in both erythrocyte and b-lymphocyte development being impaired. Additionally we may expect to see disruption to the Jak-Stat pathway, with the Jaks having an essential role in mediating the effects of hematopoietic regulators. Mice deficient in Jak1, Jak2 and Jak3 have abnormal lymphoid development and exhibit reduced numbers of functional B and T-lymphocytes (Ward et al., 2000). A further pathway which we would expect to exhibit significant disturbance in the B-cell receptor signalling pathway, a pathway which is responsible for the fate of B-cells, including survival, tolerance, proliferation, and differentiation into antibody-producing cells or memory B-cells.

As part of this investigation those pathways identified as being most likely to show disturbance (haematopoietic, B-cell signalling and Jak-Stat pathways) will be examined, identifying those genes in these pathways which are significantly differentially expressed when comparing the affected and control groups.

Why perform a pilot study?

A pilot study is a scaled down version of a full experiment, using fewer samples so that the feasibility and logistics of an experimental design can be examined, in order to identify potential problems and improve the quality and efficiency of the full experiment. Due to the limited numbers that are used within a pilot study, the data has limited use in terms of conclusions that can be drawn from the results, however it will provide vital information on the study design. Additionally, data from the pilot can be incorporated into the main study, provided that the study design is not modified (<http://www.nc3rs.org.uk/downloadaddoc.asp?id=400>). However due to the high costs associated with performing an experiment of this type, and the difficulties in obtaining suitable and high quality RNA samples, limited numbers have been used in many of the RNA-Seq experiments which have been published.

The experimental design: A pilot study to investigate global transcriptional changes in FIS-affected foals.

In an attempt to provide further understanding of this disease and enable investigation of the downstream effects of the primary mutation on gene expression, a global transcriptional study was undertaken. Due to the constraints of the sampling techniques and the number of samples available for use in this experiment, a pilot study was deemed most suitable. This would not only enable analysis of the results in terms of differential expression but would also enable a full evaluation of the sampling techniques and samples used for the study, which could be adopted for future investigations.

Tissue selection: Identifying a suitable tissue for global transcriptional analysis

Erythrocytes and lymphocytes are both derived from haematopoietic stem cells in the bone marrow. FIS is a disease of progressive anaemia and immunological dysfunction, resulting in a B-cell deficiency. Therefore, it is likely that the progenitor of both these cell types, the haematopoietic stem cell, is defective. Based on this, bone-marrow samples were collected from affected and control animals and their mRNA purified and compared to identify differentially expressed genes.

Selection of animals and sample numbers

Matching of cases and controls in any study is fundamental to the validity and reliability of the conclusions drawn from the results. However, sampling techniques were a constraint on this investigation because the same sampling technique and subsequent sample processing method could not be performed for both affected and control animals: Samples from affected foals were from bone-marrow biopsies and in contrast, the control samples were from bone-marrow aspirates. The different sampling techniques may alter the relative cell populations sampled between the affected and control animals and therefore bias the results. In addition, due to the different sample types, the samples were processed and handled differently, which again could potentially lead to biased results.

To investigate the effects of the sampling technique methodology, a bone-marrow biopsy sample from a control animal which was euthanized due to colic, was also processed as a comparison. The gene expression of this animal was compared to that of the affected and control animals, providing a guide to any inherent bias. However, this animal was suffering from colic, so undoubtedly an inflammatory response would have been initiated. Therefore any conclusions drawn from this animal have limited use.

A further limitation on this investigation was the number of samples which could be analysed, due to the high costs of a RNA-Seq experiment. Consequently this investigation was limited to a total of seven samples (split between cases and controls) as this was the capacity of one run of the sequencing equipment.

Aims and objectives of the RNA-Seq investigation

Aims:

1. To evaluate the experimental design, comparing the sampling techniques, extraction methods and results. These conclusions will provide valuable information for future studies as to the most suitable sampling technique and RNA isolation methodology.
2. To identify significantly differently expressed genes in bone-marrow tissue between FIS-affected and control animals.
3. To perform networking analysis to identify gene pathways and biological functions most affected by FIS.

Objectives:

1. Recruit animals that can be used as healthy controls. This will involve the application of a Home Office Licence, so that the sampling procedure can be carried out to obtain bone-marrow samples from these individuals.
2. Obtain bone-marrow biopsies from animals which are euthanized as they are suspected as being FIS-affected.

3. Extraction of total RNA from the affected and control animals. These samples will then be transferred to a processing laboratory for preparation of cDNA libraries and RNA-sequencing.
4. Compare the gene expression levels of the cases and control samples to identify those genes which are most significantly differentially expressed.
5. Examine the haematopoietic, B-cell receptor and Jak-Stat pathway, those pathways which are most highly associated with the FIS phenotype, for significant gene disturbances.
6. Finally, networking analysis will be performed to identify biological functions and pathways which are most affected by the FIS phenotype. Networking analysis will be performed using Ingenuity Pathways Analysis (IPA) software (Ingenuity Systems, www.ingenuity.com).

7.2 Materials and Methods

Animals and sample collection

This investigation adopted a ‘case-control’ approach comparing transcriptional levels in the affected group to a control group. At the time samples were collected for this investigation the causal mutation had not yet been identified and so controls were not selected based on genotype. It was however later identified that one of the controls was a carrier of FIS, although this did not alter the analysis as case-control studies are based only on phenotypic information.

FIS-affected animals

Bone-marrow biopsies were collected immediately post-mortem from the femur of four FIS-affected Fell Pony foals, aged 4 – 10 weeks (appendix 1). This was performed by the author and P. May, a veterinary surgeon in private practice at Newbiggen, Pentrith, The femur was sawn open approximately 4 cm from the head and bone-marrow removed using a sterile scalpel. The biopsy was then cut into 5 × 5 mm pieces, immersed in five times volume RNAlater solution (Ambion, USA), which acts to stabilise and protect the cellular RNA, and then stored at +5 °C for 24 hrs. The samples were then removed from the RNAlater solution before archiving at -80 °C until required for RNA extraction.

These FIS-affected samples were collected prior to the discovery of the highly associated *SLC5A3* mutation, so diagnosis could not be performed at the time using PCR. Rather, disease status was confirmed post-mortem based on gross pathological findings, histological and haematological analysis. Later, disease status was definitively confirmed; all four foals were homozygous for the *SLC5A3* mutation.

Control samples

A 4 ml bone marrow sample was aspirated using standard medical practices, into a syringe containing the anticoagulant ethylenediamine tetra-acetic acid (EDTA) by Prof. D. Knottenbelt under a Home Office Licence (PPL Number: 80/1916) from

seven healthy Fell Pony foals (aged 10 – 14 weeks). The samples were then handled in accordance with the PAXgene Bone Marrow RNA Tube Product Circular (<http://www.qiagen.com/products/rnastabilizationpurification/paxgenernasystem/paxgenebonemarrowrnakit.aspx#Tabs=t2>) (Qiagen, UK).

To assess the samples for cell populations, a fresh bone-marrow smear was prepared from all of the samples for microscopic examination. The bone-marrow smears were examined by S. Putwain, resident in clinical pathology, Cambridge University, for complete maturation sequence identification. Also examined were flecks of bone-marrow tissue, to confirm the presence of all expected cell types (Appendix 8).

An additional control sample was collected from a six week-old Fell Pony foal which was euthanized due to colic (appendix 1). A bone-marrow biopsy was collected immediately post-mortem using the same protocol as the FIS-affected samples.

All control samples were subsequently screened for the *SLC5A3* mutation to assess FIS-carrier status, using the methodology described in chapter six.

Total RNA extraction from bone-marrow samples

Prior to RNA isolation, the work bench, pipettors and all required equipment were RNAase decontaminated using RNaseZap wipes (Ambion, USA) according to the manufacturer's instructions.

Total RNA isolation from bone-marrow biopsies

Total RNA was isolated from the bone-marrow biopsies using the RNAqueous kit (Ambion, USA), in accordance with the manufacturer's instructions. Samples were prepared by homogenising 50 mg of frozen bone-marrow tissue, which was sliced from the frozen biopsies using a scalpel blade, in 200 µl of lysis-binding solution (supplied in the RNAqueous kit) in 500 µl RNase-free microfuge tubes (Ambion, USA) using a disposable homogeniser. The lysate was then passed through a 25 gauge syringe needle several times until the tissue was liquefied, and RNA isolated by following the standard RNAqueous kit protocol. RNA was eluted in two steps with the provided elution buffer, firstly using 30 µl and then an additional 20 µl. A

10 µl aliquot was taken for quantification and quality analysis and the remaining sample was stored at -70 °C until required.

Total RNA isolation from bone-marrow aspirates

Initially the samples were equilibrated to room temperature, carefully inverted ten times, and then kept at room temperature for two hours prior to transferring into RNase-free 15 ml conical tubes (Ambion, USA). Total RNA was then isolated using the PAXgene Bone Marrow RNA Kit (Qiagen, UK), in accordance with the manufacturer's instructions. A 10 µl aliquot was taken for quantification and quality analysis and the remaining sample was stored at -70 °C until required.

Quantification and assessment of the RNA quality

RNA integrity and quantification were assessed using an Agilent 2100 bioanalyzer (Agilent, USA) with the RNA 6000 Nano Kit (Agilent, USA), in accordance with the manufacturer's instructions. RNA integrity was determined from the RNA integrity number (RIN score) (Schroeder et al., 2006), which is automatically calculated by the Agilent 2100 bioanalyzer from the 18S to 28S ribosomal band ratio. Only those samples with an RIN score >9 were used for the RNA-Seq investigation (Wilhelm et al., 2010).

RNA-Sequencing using the Illumina Genome Analyser II System

Isolation of mRNA from total RNA, preparation of cDNA libraries and the sequencing of the cDNA libraries, was outsourced to Source BioScience, Nottingham. Five micrograms total RNA in a 30µl volume was supplied for each sample. The samples were processed and sequenced on the Illumina Genome Analyser II-x (GAIIx) machine, obtaining single end 38 bp reads. The GAIIx was run for 38 cycles.

An overview of sample preparation for RNA-Sequencing

Detailed here is an overview of the RNA-Seq protocol (Wilhelm and Landry, 2009, Nagalakshmi et al., 2010). Initially, the ribosomal RNA (rRNA), which forms the vast majority of RNA (>90%) present in cells, and which is essentially

uninformative, must first be removed. In this experiment, rRNA was removed using polyA enrichment. The polyA enrichment step is the most commonly used method used to enrich for mRNA. This method uses beads which have long oligodT stretches on them that bind to the polyA tails that are present on most mRNA molecules, thus enabling the selection of non-ribosomal transcripts.

Following enrichment of mRNA, reverse transcription is primed using either random primers or oligo dT primers. Random primers avoid 3' end bias, whereas oligo dT primers can result in a bias towards the 3' end due to an underrepresentation of the 5' end in longer transcripts. However, the advantage of oligo dT primers is that the majority of the cDNA produced is from polyadenylated mRNA and so a higher proportion of sequence obtained is informative.

Finally, the first-strand cDNA is converted to double-stranded cDNA. This is prepared by removing the RNA from the DNA-RNA hybrid and synthesising a replacement strand to yield double-stranded cDNA. The double-stranded cDNA is then fragmented and ligated to Illumina adapters for subsequent amplification and sequencing.

RNA-Sequencing using the Illumina Genome Analyser II platform

The RNA-Seq protocol is based entirely on the same principles as DNA sequencing using a next-generation sequencing platform. Here, the Illumina Genome Analyser II was used to sequence the cDNA libraries, a sequencing technology which is based on the sequencing-by-synthesis principle (Fig 7.1). The sequencing-by-synthesis methodology is based on a three step protocol: Library preparation, cluster preparation, followed by sequencing of the cluster libraries. Firstly, the sample is fragmented and adaptor oligos which are complementary to those on the array are ligated to the ends of the fragments. These adaptor oligo-ligated fragments are then applied to one channel of an 8-channel glass flow cell, and hybridised to the lawn of complementary oligos which are bound to the surface of the glass array. The molecules subsequently bend over, forming bridge-like structures, forming a template for complementary strand synthesis. Each template is then sequenced in parallel, where each cluster is supplied with four different fluorescently labelled nucleotides that have their 3'-OH group chemically inactivated to ensure that only a single base can be incorporated during one cycle. The incorporated nucleotide is then

deduced by exciting the bound fluorescent labels, the signal recorded and the 3' end unblocked in preparation for the following cycle (Mardis, 2008).

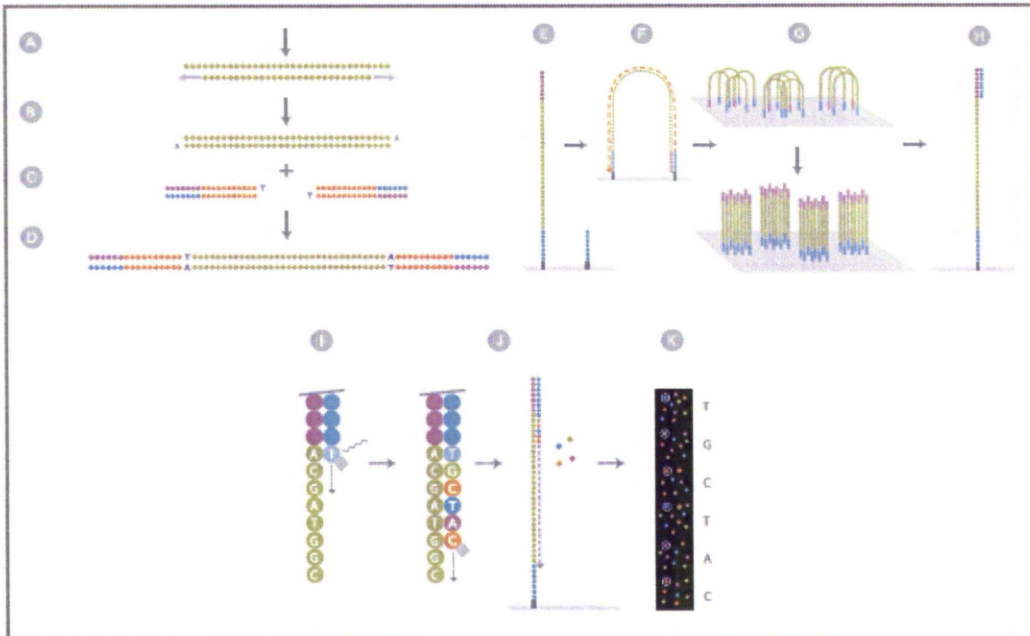


Figure 7.1: An overview of the Illumina Genome Analyser workflow. Initially the sample is fragmented (a) and the end repaired (b), ready for adaptors to be ligated to the repaired ends (c). The ligated fragments are selected (d) and attached to the flow cell (e) for bridge amplification (f). Clusters are then generated and the sequencing primers annealed to the clusters ready for sequencing (g and h). A single nucleotide is then incorporated into the extending fragment (i and j) and the incorporated nucleotide is deduced by the image reader (k). Step (i, j and k) is then repeated, extending the strand and deducing the sequence.

Analysis of RNA-Seq data: Detecting differentially expressed genes

Analysis of the RNA-Seq data was conducted by G. Blackshield at Source BioScience, Nottingham:

Initially, RNA-Seq data was analysed using the Illumina Genome Analyser pipeline software (version 1.5.1), to undertake image analysis and base calling. The data was then analysed using Generation of Recursive Analyses Linked by Dependency software (GERALD). Using this program, an initial alignment to the reference

sequence was performed, to assess the quality of the data and filter out non-unique alignments and overlapping clusters.

Full analysis was then carried out on the filtered data, using a pipeline of interrelated programs; Bowtie (<http://bowtie-bio.sourceforge.net/index.shtml>), TopHat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2010). Using Bowtie to map the reads, TopHat (version 1.1.0) was used to align the filtered RNA-Seq reads to the reference genome, in order to identify alternative splice junctions. TopHat creates a number of files, including one with a complete list of alignments (SAM file) and a UCSC BedGraph file (WIG file) which can be uploaded into the UCSC browser to visualise depth of coverage, and splice variants at each position. However, due to the limited dataset in this investigation examining for alternative splice variants would be crude. Examination of the BedGraph on the UCSC browser would only simply enable truncation of genes to be identified rather than splice variants as identification of splice variants using global transcriptional data is extremely complex, requiring high computational power and high read-depth (Pan et al., 2008). All samples had the read depth and splice variants examined for the four genes in the FIS 375,043 Kb critical interval. The BedGraph file was visually examined to assess any differences in read depth and splice variants between the control and cases. The alignment phase was then followed by analysis using Cufflinks (version 0.8.2), which assembles the transcripts and estimates their abundance.

Following this, differential expression tests were carried out between the affected and control samples using DESeq from the R package (Anders and Huber, 2010). Initially, the files generated from TopHat and Cufflinks, which specify gene level coordinates and read counts, were loaded into the DESeq Module and each sample was identified as control or affected. Prior to performing the analysis, sample clustering was assessed on a “heatmap”. This identified outliers which did not cluster within their expected phenotypic group and should therefore be excluded from further analysis. Once any outliers had been removed, the estimated variance across the groups was calculated. This was required because a core assumption of DESeq is that the mean is a good predictor of the variance, i.e. that genes with a similar expression level also have similar variance across biological replicates. Therefore, prior to performing any differential tests, an estimate of variance for each condition and variance from the mean was calculated. This estimation was done by calculating,

for each gene, the sample mean and variance across biological replicates and then fitting a curve to this data. This data was then plotted to show the estimated variances and to determine whether the mean was a reliable predictor of the variance. It should be noted that only where there are an appreciable number of biological replicates can the variance estimates be expected to be accurate. It is instructive to observe at which count level the biological signal starts to dominate the background noise. At low counts, where background noise dominates, greater sequencing depth (larger library size) will improve the signal-to-noise ratio, while for high counts, where the biological noise dominates, only additional biological replicates will help. An additional check to further confirm similarity of gene expression across biological replicates was also performed, to assess the per-gene variance, identifying if the base variance followed the empirical variance.

Once the variance-mean dependence was estimated and verified, the two groups were compared to identify differentially expressed genes. This is performed by contrasting conditions in a pair-wise manner. For each gene, the fold change was calculated from the mean expression level for both conditions, along with the logarithm (base 2) of the fold change and the *P*-value for the statistical significance of this change. Finally, the adjusted *P*-value was calculated, by adjusting for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), which controls false discovery rate (FDR, set at 0.1%). A FDR (Farcomeni, 2008) of 0.1% was used for this investigation and therefore we would expect 1/1000th of the tests to be type 1 errors (i.e. false positives), a reasonable error rate for this dataset (G. Blackshields, personal communication). Graphical visual representation of differentially expressed genes was then plotted, displaying the logarithm (base 2) of the fold change against the base means.

Using KEGG Mapper, freely available web based software (http://www.genome.jp/kegg/tool/map_pathway2.html) the haematopoietic, B-cell receptor and Jak-Stat pathways will be examined for disturbances. Genes which are significantly differentially expressed in these pathways will be highlighted through the use of colour: Genes which are up-regulated in the affected group with respect to the control group will be highlighted in red and those shown in green are significantly down-regulated. Producing these images will provide an overview of these pathways and the extent of the gene disruption within this pathway.

Analysis of molecular interactions

The quantity of data provided by RNA-Seq enables close analysis of how changes in gene expression may have direct or indirect effects on the normal physiology of an organism. Consequently, there are several ways in which this data can now be analysed and interpreted. Pathway analysis examines a group of genes which rely on one another to perform a specific biological function. Within a genetic pathway, the disruption of a single gene, for example by a mutation, will affect the downstream pathway. Pathway analysis can provide information on those biological processes which are significantly disrupted by the disease, and which in turn can provide an insight and understanding of the clinical presentation of a disease.

Networking analysis explores the complex network of interacting cellular components, depicting how single molecules interact with one another, by both direct and indirect interactions. Gene networking analysis also provides a guide to the physiological state of an individual, as most phenotypes arise from the response of a large number of genes with coordinating activity. Therefore, single mutations can affect multiple cellular functions, which in turn give rise to diverse phenotypes. Gene networks are constructed by specialist software based on known gene functions and interactions. Usually, the outputs from these gene networks are displayed as directional graphs, showing positive interactions with an arrow, and negative interactions with a bar at the end of the line.

There are many approaches and commercial software packages available for investigating pathway interaction. Here, we used a free two-week trial of a web-based program, Ingenuity Pathway Analysis, version 8.7 (IPA) (<http://www.ingenuity.com/index.html>) to examine for those pathways which are significantly altered in FIS-affected animals. IPA performs an analysis based on information from the Ingenuity knowledge base, a library which contains information on protein interactions, regulatory events and gene to phenotype information. The Ingenuity knowledge base is based on literature sources, including journal articles, review articles, textbooks and databases such as OMIM and EntrezGene. IPA analysis can identify networks that contain either “direct” and/or “indirect” relationships, an option which is specified by the user when commencing

analysis (Deighton et al., 2010). Direct interactions are defined as those where two molecules make direct physical contact with each other with no intermediate step, whereas indirect interactions are defined as a relationship between two molecules via intermediate steps. For this investigation, both direct and indirect relationships were examined, and a fold-change threshold of two specified for inclusion.

An overall score is computed for each network, a score which is derived from the *P*-values (corrected using Benjamini Hochberg), indicating the likelihood of gene expression being altered due to random chance. The network score is the negative log of this probability estimate (*P*-value), so a network score of 2 indicates that there is a 1 in 100 chance that the network is altered due to chance. In addition, IPA also computes *P*-values for biological functions, so the investigator can assess those most significantly affected by the disease.

7.3 Results

Animals and Samples: Selection of samples for RNA-Seq

Samples were selected for the RNA-Seq experiment based on histology reports and the quality of the RNA extracted. A total of four FIS-affected bone marrow samples were selected along with three control bone marrows, of which two (one FIS-clear and one FIS-carrier) were from bone-marrow aspirates. The third control was from a bone-marrow biopsy which was taken from a carrier animal euthanized due to colic. It should be noted that the control samples were not selected based on their phenotype or genotype. This is because at the time the samples were selected, the genotyping test described in chapter 6 was not yet available. It was only after the samples were selected and the RNA-Seq experiment conducted, that the carrier status of these individuals was confirmed.

Histology Reports for the bone-marrow aspirate smears

A bone-marrow smear was prepared for each of the seven bone-marrow samples that were aspirated from healthy control animals, and sent for cytological examination. This was to confirm the presence of expected cell types typical of bone marrow. Of the seven, five were reported to be unrepresentative of bone-marrow; no flecks of bone-marrow were observed, no megakaryocytes or blast cells were seen, there was marked dilution from peripheral blood, and many of the segmented neutrophils were derived from peripheral blood rather than bone marrow. Based on the report (Appendix 8), these five samples were excluded from the RNA-Seq experiment as they did not provide a true representation of the bone-marrow tissue and so would not be suitable as a comparable tissue to the bone-marrow biopsies from the FIS-affected ponies.

The bone-marrow aspirate from sample 20_01 (FIS-clear) had no observed bone-marrow flecks, but rubricytes and metarubricytes were seen. Again, there were no megakaryocytes observed but there were neutrophils and the myeloid to erythroid ratio was approximately 1:1.8 (normal ratio 2:1 to 4:1). The bone-marrow aspirate

from 21_01 (FIS-carrier) had the highest cellularity and although no bone-marrow flecks were observed, a single megakaryocyte was seen. In addition there were occasional metamyelocytes, segmented neutrophils and eosinophils. Occasional lymphocytes, metarubricytes and rubricytes were seen, although there was no complete maturation sequence of either line observed. Based on the report, only these two control bone marrow aspirates were selected for RNA-Seq analysis. However, due to the underrepresentation of bone-marrow tissue in these samples, it is likely that some of the differential expression detected will be the direct result of the different cell ratios, resulting from the sampling techniques, rather than due to disease.

Determining the quality of the RNA using a BioAnalyser

Quality and integrity of total RNA was determined using a BioAnalyser (Agilent, USA). All seven extracted samples that had been selected for the RNA-Seq experiment achieved RIN scores > 9 (Table 7.1).

Sample	20_09	21_09	04_08	01_08	03_09	04_09	06_09
Affection status	Clear	Carrier	Carrier	Affected	Affected	Affected	Affected
RIN Score	9.70	9.80	9.70	10	10	10	10
Sampling technique	Aspirate	Aspirate	Biopsy	Biopsy	Biopsy	Biopsy	Biopsy

Table 7.1: RIN scores which indicate the quality and integrity of the RNA extractions.

Mapping the reads

The seven samples were processed and sequenced on the Genome Analyser II-x machine for 38 cycles. After base-calling and purity filtering using the Illumina Genome Analyser Pipeline Software (version 1.5.1), 4,649,547 Kb of sequencing

data was obtained from the seven samples. To initially assess read quality, Generation of Recursive Analyses Linked by Dependency software (GERALD) was used to align the reads to the *Equus caballus* reference genome (EquCab2 assembly), uniquely mapping 3,263,372 Kb (70.19%) of the filtered reads to the reference genome (table 7.2).

Sample	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	% Align (PF)
06_09	653145	229659	143234	71.32
03_09	674644	228563	147948	66.33
01_08	613825	241494	134611	71.13
04_09	660104	236904	144760	71.40
04_08	704554	232499	154508	71.50
20_09	673461	236463	147689	70.65
21_09	669814	225626	146889	69.06

Table 7.2: Summary of the data quality metrics and basic alignments for quality filtering. Clusters (raw) is the number of clusters that were detected by the image analyser; Illumina recommends that this exceeds 20,000. Clusters (PF) is the number of clusters which passed the purity filter. Percentage align (PF) is the percentage of filtered reads which mapped uniquely to the reference genome.

Identifying differentially expressed genes

Advanced analysis of the RNA-Seq data was performed using a pipeline of interrelated programs. Initially, TopHat, a program which uses Bowtie to map the reads, was used to align the RNA-Seq reads to the reference genome, increasing the number of aligned reads in all samples compared to that observed from mapping with the Illumina Genome Analyser Pipeline Software (Table 7.3).

Sample	Number of reads aligned (Bowtie)	% of reads aligned
06_09	13,036,657	75.87
03_09	12,390,468	69.82
01_08	12,243,104	75.85
04_09	13,154,441	75.81
04_08	14,253,334	76.98
20_09	13,171,311	74.39
21_09	12,808,425	72.74

Table 7.3: *Number of reads aligned using TopHat. All samples had an increased number of reads that mapped, compared to the number of reads which aligned using the Illumina Genome Analyser Pipeline Software.*

After completing the alignment, TopHat produces WIG files for each sample which can be uploaded into the UCSC genome browser to investigate alternative splicing and truncation of genes. The alignment of the FIS 375,043 Kb critical interval for each sample was viewed in the UCSC browser to investigate alternative transcripts (alignments shown in Appendix 9). These alignments provided no evidence of alternative splice variants or gene truncation within the critical interval; there was an excellent match between the alignments of all seven samples.

Cufflinks was then used to generate the input files for DESeq analysis by assembling the transcripts and estimating the read counts for all of the samples. The files were then loaded into the DESeq module to identify differentially expressed genes. Initially, sample clustering was performed to provide an overall visual representation of how similar or dissimilar the two sample groups were (Fig 7.2). Close grouping of the samples shows that they had an overall similar gene expression pattern, so it was expected that the three controls would group together and the four affected samples

would group together, showing a distinct separation of the two groups. The heatmap clearly showed that two of the control samples grouped closely together as did the majority of the affected samples. However, the control sample which was taken post-mortem by biopsy (sample 04_08), clustered with the affected group. This suggested that sampling technique, sample processing method and/or disease pathology had caused this sample to cluster with the affected samples. This sample now formed an additional group ‘unique’ and because of this anomaly, this sample was excluded from networking analysis, including those pathways drawn using KEGG Mapper.

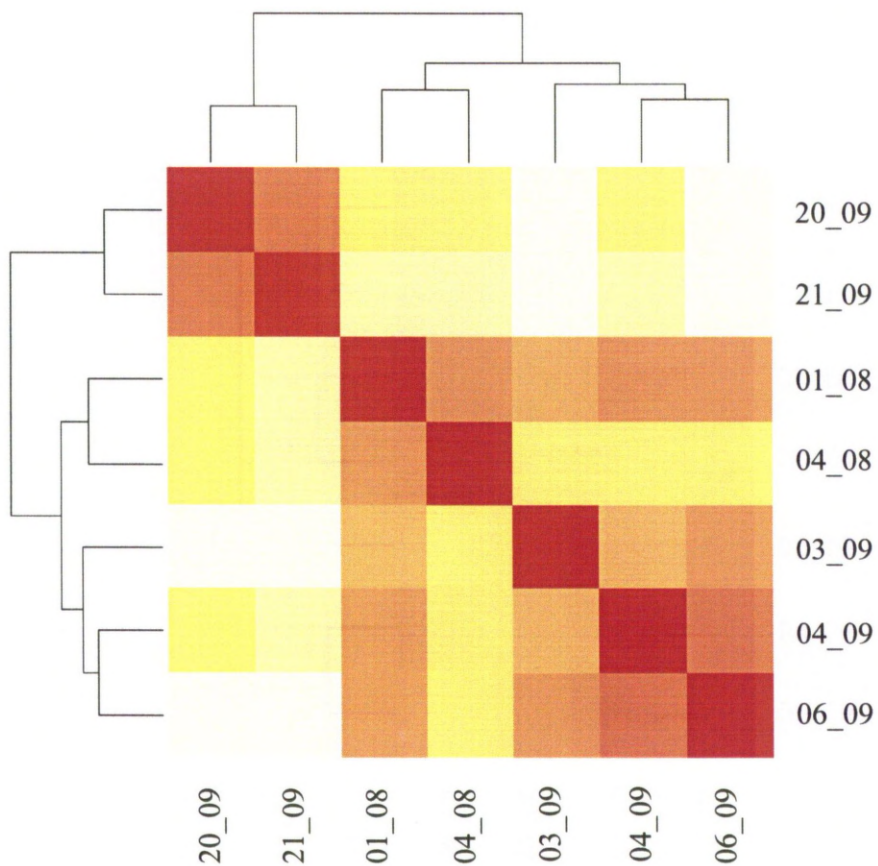


Figure 7.2: Heatmap showing the clustering of the individual. The Heatmap provides graphical representation of the data as a two-dimensional table. The colours represent pairwise comparisons between two samples with largest values (samples showing very similar levels of gene expression) being displayed in red (hot), the smallest values in blue (cool), and intermediate values in shades of orange. From the image, it is clear that two of the control samples (20_09 and 21_09) group closely

together. The majority of the affected samples also group together (01_08, 03_09, 04_09 and 06_09)), but the control sample, 04_08), nestles within the affected group.

To assess gene expression levels across the biological replicates, the variance from the mean was calculated. For each gene, the sample mean and variance across biological replicates were calculated and then a curve was fitted to this data (Fig 7.3). This illustrates the variation of the squared coefficient of variation (SCV; the ratio of the variance at base level to the square of the base mean) against the base mean. The solid lines are the SCV for the raw variances (noise due to biological replication), with one solid line per replicated condition. On top of the variance there is shot noise (variance inherent to the process of counting reads). The amount of shot noise depends on the size factor, and hence, for each sample, a dotted line for each condition is plotted above the solid line. The dotted line is the base variance, showing the full variance, scaled down to base level by the size factors. The vertical distance between the solid and dotted lines is equal to the shot noise and the solid black line depicts the density estimate of the base means. From the pattern observed in Fig 7.3, it is clear that the mean is a good predictor of the variance across replicates and that control sample 04_08 had highly similar expression levels to the affected group.

A further check to assess that genes with a similar expression level also have similar variance across replicates is to assess whether the base variance follows the empirical variance. As can be seen in Fig 7.4, the single-gene estimates for the control and affected groups follow the line of best fit very well, again providing support for replicates having similar expression levels.

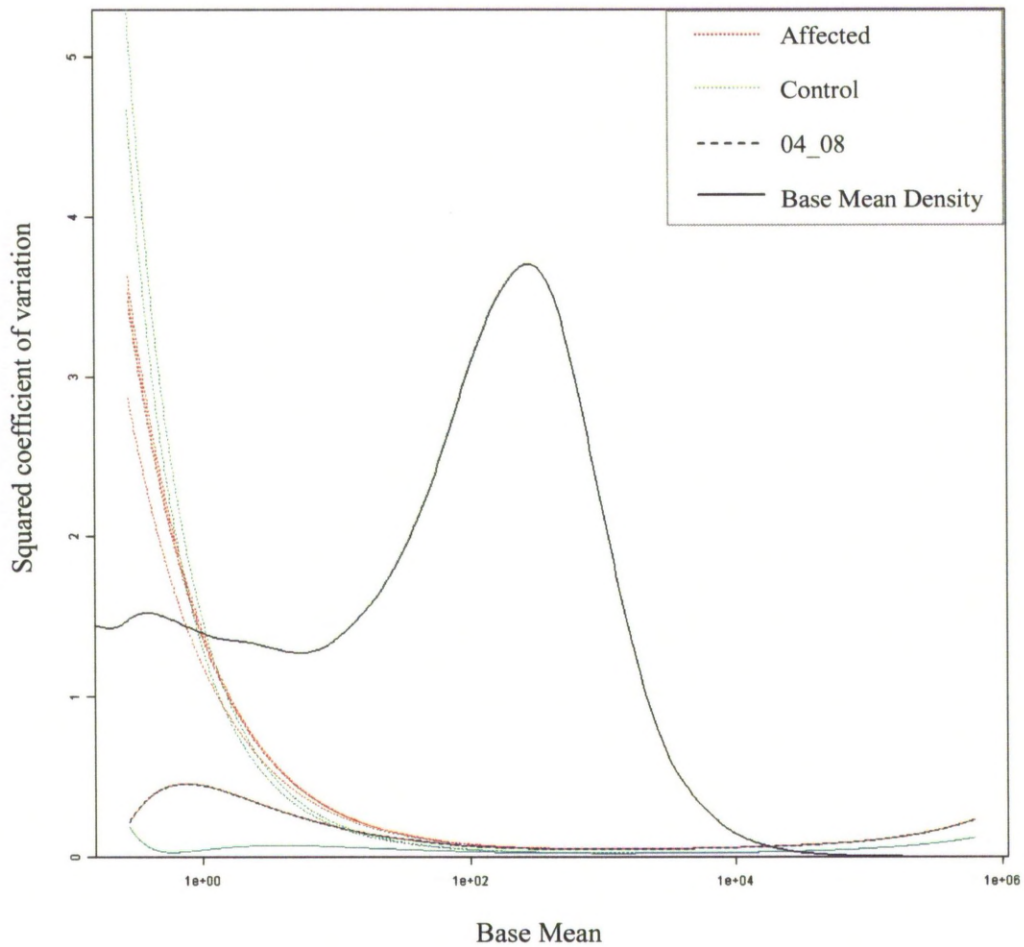


Figure 7.3: Estimated variances as squared coefficients of variation. From the pattern visible, it is apparent that the mean is a good predictor of the variance across replicates. Sample 04_08 (dashed black line) has also been shown here to have highly similar expression levels to the affected group.

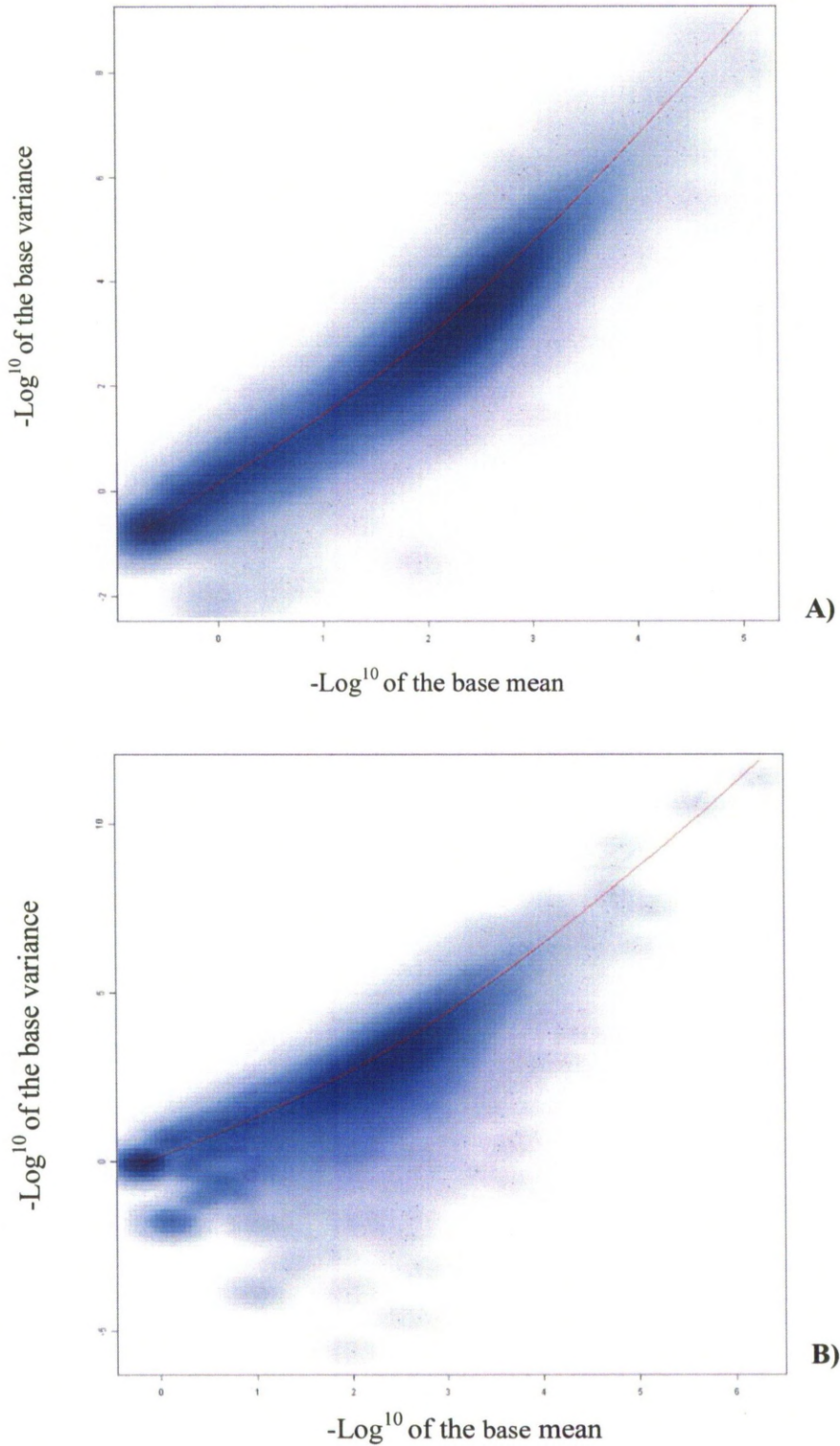


Figure 7.4: Per-gene estimate of variance for (a) Affected and (b) Control groups. The blue cloud summarises all per-gene estimates. Darker shaded blue represents increased density of genes, shaded through to grey (few genes). The red line indicates the best fit from the logistical regression. Both groups follow the logical

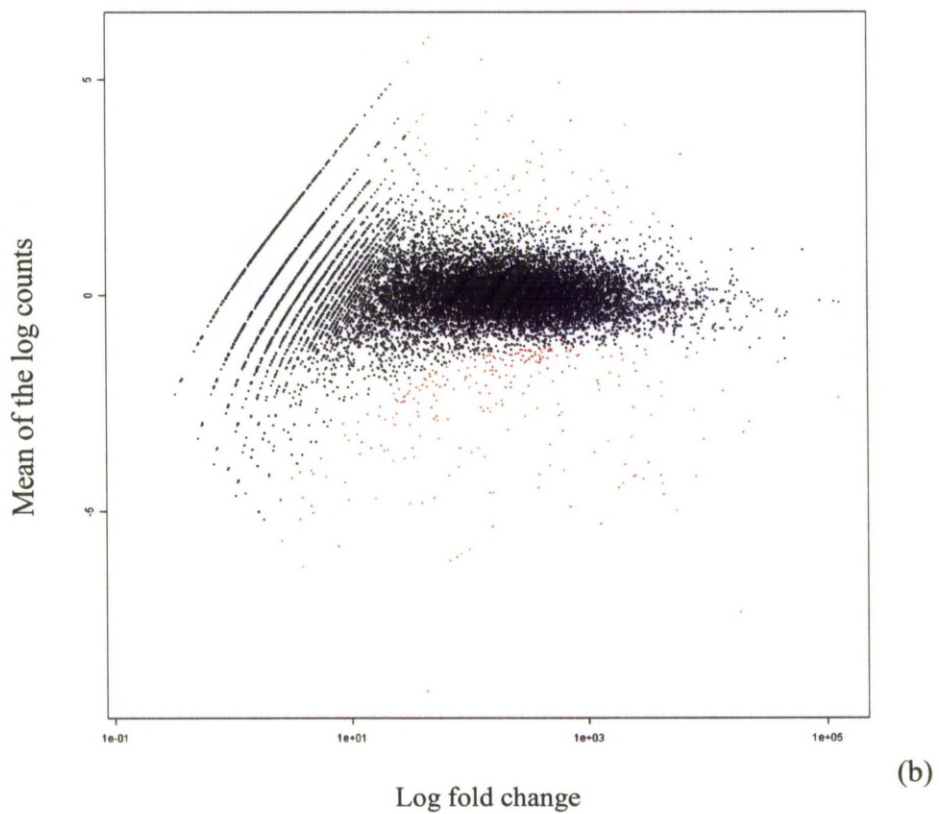
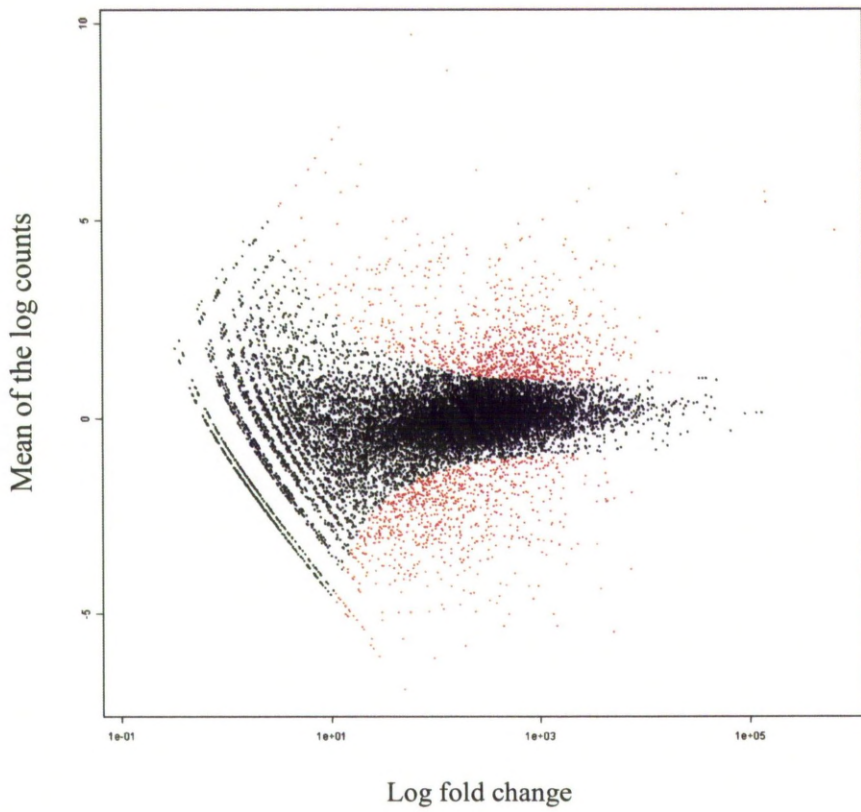
regression best fit line well (darker shading along the line of best fit compared to the surrounding area), considering how few samples were used for the experiment.

Now that the variance-mean dependence had been estimated and verified, the two groups could be compared to identify genes differentially expressed at a 0.1% false discovery rate. Three pair-wise comparisons were undertaken, comparing the mean expression level for each gene, calculating the fold change, and its statistical significance. The three pair-wise comparisons were (Fig 7.5): a. Affected Vs Control, b. Affected Vs Unique and c. Unique Vs Control. The two comparisons using the unique group (sample 04_08) were conducted to identify how similar/dissimilar this sample was to the affected and control groups. The affected and control pair-wise comparison identified 1895 differentially expressed genes, of which 854 were up-regulated in the affected group and 1041 were down-regulated in the affected group. Of those genes which were up-regulated in the affected group, 18 were not expressed at all in the control group (Table 7.4). Those top twenty differentially expressed genes, both up and down regulated are shown in Tables 7.5 and 7.6. The affected and unique pair-wise comparison identified 187 differentially expressed genes, of which 41 were up-regulated in the affected group and 146 were down-regulated in the affected group. The unique and control pair-wise comparison identified 770 differentially expressed genes, of which 233 were up-regulated in the control group and 537 were down-regulated in the control group.

Gene ID	Functional details
<i>NRXN1</i>	Forms part of the receptor for the nervous system of vertebrates.
<i>MGP</i>	Acts as an inhibitor to bone formation.
<i>COL6A6</i>	Acts as an anchoring meshwork with mutations being associated with cell spreading.
<i>ARHGAP20</i>	Unknown
<i>SLC24A4</i>	key signal for the regulation of store operated channels, to allow an influx of Ca ²⁺ into non excitable smooth muscle
<i>HHIP</i>	Important for a wide range of developmental processes. Mutations are associated with human height.
<i>CDH7</i>	Mutations are highly associated with CHARGE syndrome, a condition which results in multiple embryonic malformations.
<i>KCNT2</i>	Member of the calcium-activated potassium channel protein family
<i>ANO1</i>	Mutations are associated with gastroparesis, a condition which results in slow emptying of the gut.

<i>C10orf81</i>	Unknown
<i>LBP</i>	Associated with Chrohns disease. Protein encoded for by this gene is involved in the acute-phase immunologic response to gram-negative bacterial infections and plays an important role in the LBP dependent monocyte response.
<i>SIX2</i>	Involved in the development of the limbs and eyes. Mutations are associated with abnormal kidney development.
<i>CHRDL2</i>	Highly expressed in many human tissues and is particularly abundant in the uterus. CHRDL2 plays an important role in the differentiation of myoblasts and osteoblasts
<i>Q9N0F0-HORSE</i>	Unknown
<i>RBP5</i>	Predicted to have a role in the control of the cell cycle. Down regulation of RBP5 is associated with aggressive hepatocellular carcinomas
<i>A6P3C6-HORSE</i>	Unknown
<i>STK32C</i>	The protein encoded by this gene is a member of the serine/threonine protein kinase family, with the specific function of this kinase being unknown.
<i>ENSECAG00000015356</i>	Unknown

Table 7.4: *Functional details of those genes which showed up-regulated expression in FIS-affected animals compared to the controls. All of these genes were not expressed at all in the control samples at the time of sampling.*



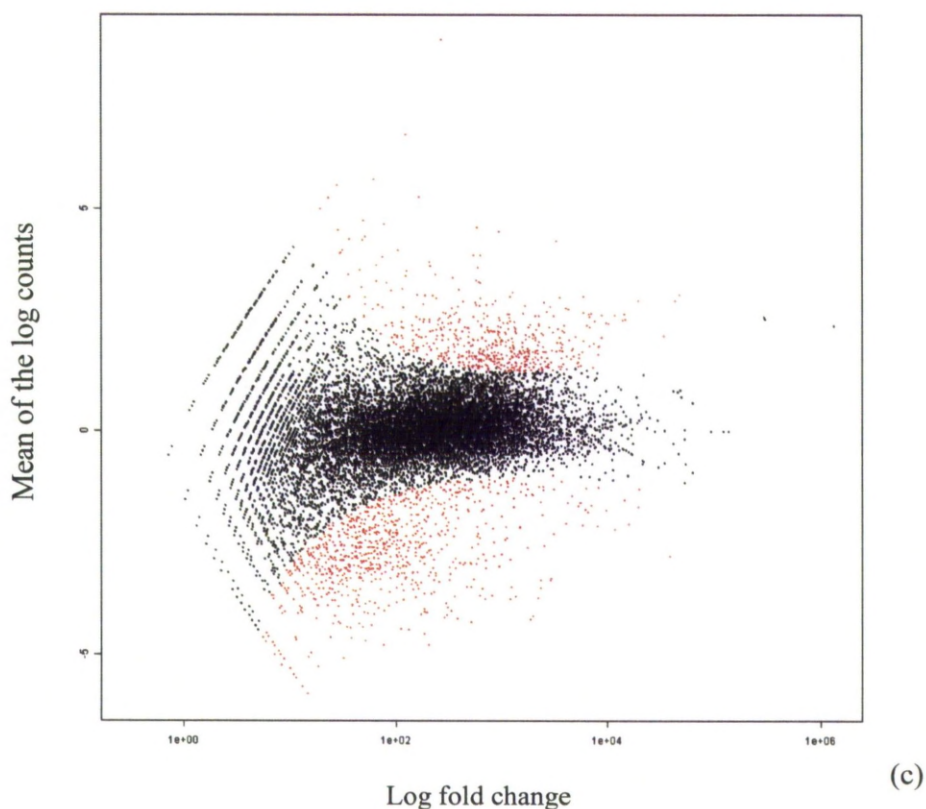


Figure 7.5: Summary of the fold change between the three pair wise comparisons which were performed. The red dots represent the genes that were classified as being statistically differentially expressed at a 0.1% false discovery rate. A: Affected Vs Control identified 1895 genes as being differentially expressed. B: Affected Vs Unique identified 187 genes as being differentially expressed. C: Unique Vs Control identified 233 genes as being differentially expressed.

Gene ID	Expression level (counts)		log ₂ of Fold Change	P-value	P-value (corrected)	Gene Information
	Affected Mean	Control Mean				
	Base mean (Affected)	Base Mean (Control)				
<i>NRXN1</i>	46.63586445	0	-Inf	2.01E-15	4.93E-14	neurexin 1
<i>MGP</i>	32.36642908	0	-Inf	9.95E-12	1.82E-10	matrix Gla protein
<i>COL6A6</i>	27.54397397	0	-Inf	2.03E-10	3.38E-09	collagen, type VI, alpha 6
<i>ARHGAP20</i>	27.27813854	0	-Inf	1.8E-10	3.56E-09	Rho GTPase activating protein 20
<i>SLC24A4</i>	27.2617638	0	-Inf	1.98E-10	3.26E-09	solute carrier family 24 (sodium/potassium/calcium exchanger), member 4
<i>HHP</i>	26.49823567	0	-Inf	2.88E-10	4.65E-09	hedghog interacting protein
<i>CDH7</i>	24.84597788	0	-Inf	1.43E-09	2.18E-08	cadherin 7, type 2
<i>KCNT2</i>	21.94109652	0	-Inf	8.68E-09	1.23E-07	potassium channel, subfamily T, member 2
<i>ANO1</i>	20.47310474	0	-Inf	2.22E-08	2.99E-07	anoctamin 1, calcium activated chloride channel
<i>C10orf81</i>	19.02028849	0	-Inf	6.16E-08	7.87E-07	PH domain-containing protein C10orf81 (Epididymis luminal protein 185)
<i>LBP</i>	17.95794652	0	-Inf	2.67E-07	3.17E-06	lipopolysaccharide binding protein
<i>SLX2</i>	17.18837691	0	-Inf	4.49E-07	5.18E-06	SIX homeobox 2
<i>CHRD12</i>	16.71548013	0	-Inf	4.19E-07	4.85E-06	chordin-like 2
<i>QSOX10-HORSE</i>	15.65026979	0	-Inf	9.53E-07	1.06E-05	A.drenomedullin Fragment
<i>RBP5</i>	13.66151623	0	-Inf	6.89E-06	6.81E-05	retinol binding protein 5, cellular
<i>AGP3C6-HORSE</i>	13.29564472	0	-Inf	7.40E-06	7.27E-05	Potassium large conductance calcium-activated channel, subfamily M, alpha member 1
<i>STK32C</i>	13.22374825	0	-Inf	6.36E-06	6.33E-05	serine/threonine kinase 32C
<i>ENSEC4G00000015356</i>	13.07207445	0	-Inf	9.46E-06	9.14E-05	no information available
<i>OLFML2A</i>	76.10144276	0.62579	-6.92610	3.78E-22	1.44E-20	olfactomedin-like 2A
<i>CNTN2</i>	144.5409642	2.05362	-6.13716	1.22E-34	9.32E-33	contactin 2 (axonal)
<i>SLC5A3</i>	168.1915715	33.86304	-2.31232	1.74E-15	4.29E-14	solute carrier family 5 (sodium/myo-inositol cotransporter), member 3

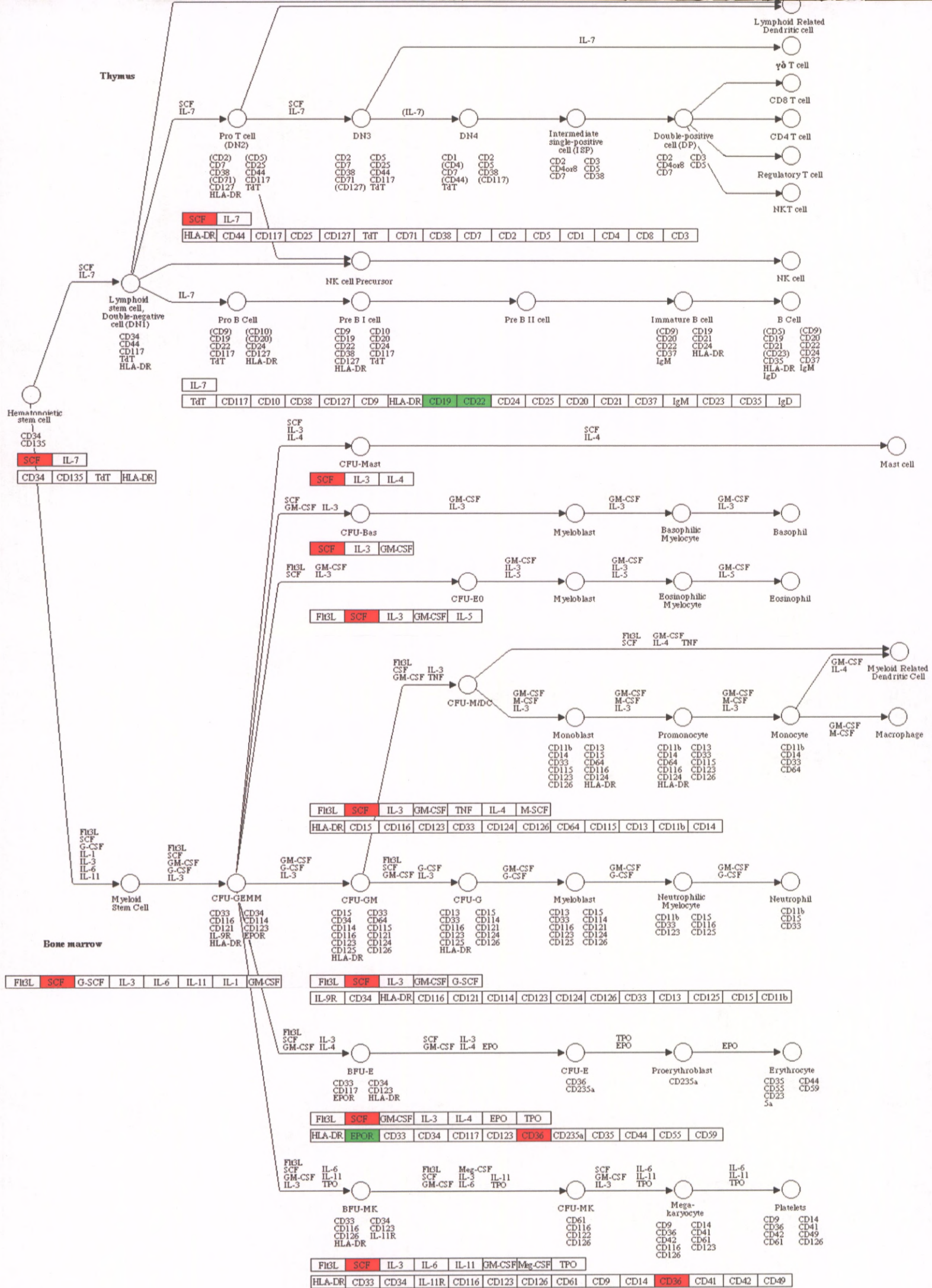
Table 7.5: The top twenty genes which were identified as up-regulated in the affected animals compared to the control group. Also shown is the expression level of *SLC5A3*, which is the gene that a highly associated mutation was identified in (highlighted in blue).

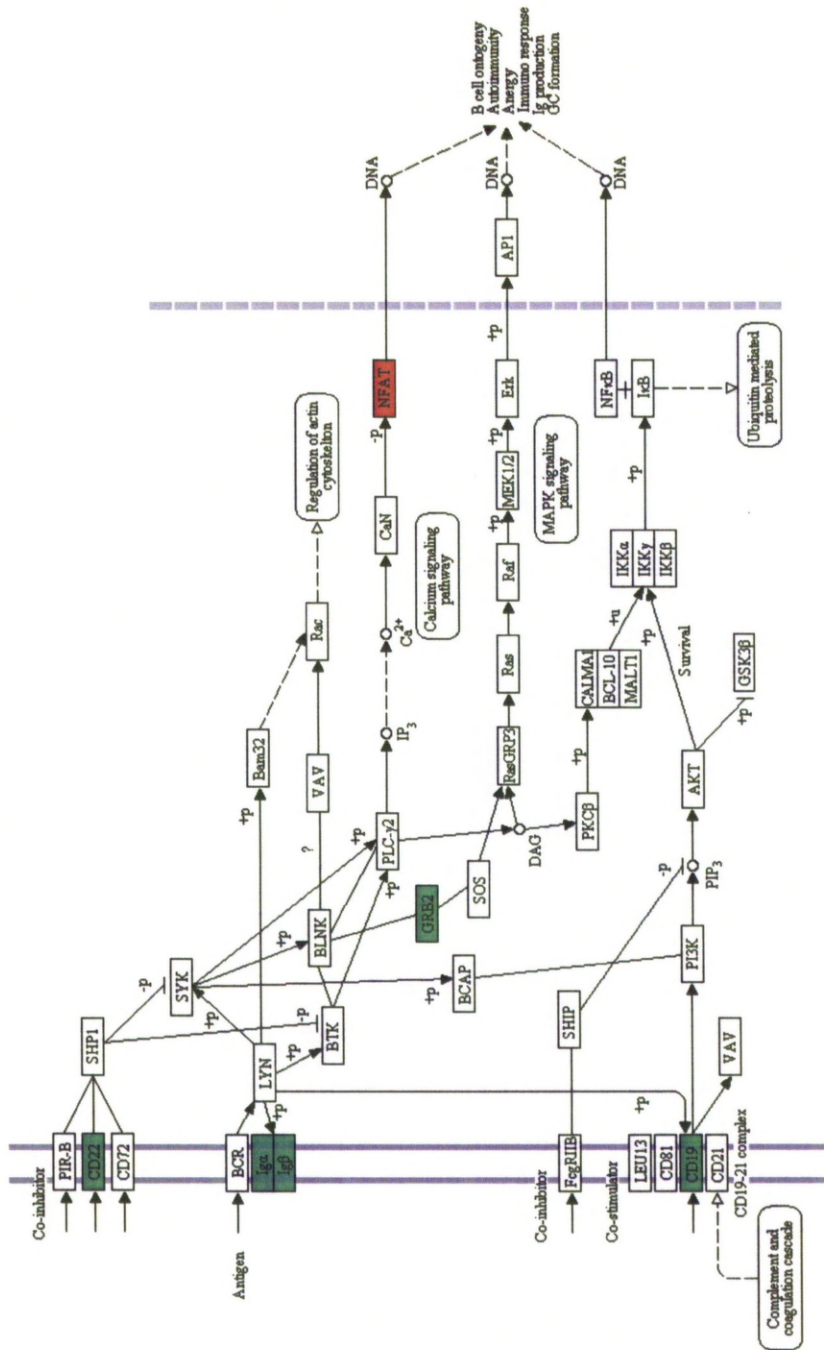
Gene ID	Expression level (counts)		Fold Change	log ₂ of Fold Change	P-value	P-value (corrected)	Gene Information
	Affected Mean	Control Mean					
	Base mean (Affected)	Base Mean (Control)					
SAMD7	0	20.53624	Inf	Inf	1.43E-11	2.59E-10	sterile alpha motif domain containing 7
ENSECAG00000000528	0	15.53883	Inf	Inf	1.50E-09	2.27E-08	no information available
ENSECAG00000012207	0.231566549	9.91562	42.81976	5.42020	3.80E-06	3.88E-05	no information available
FBAT-HORSE	9683.500711	418470.57642	43.21480	5.43345	4.88E-123	5.23E-120	Hemoglobin subunit theta-1
FAM46C	133.052067	6755.66296	44.13964	5.46400	3.04E-138	6.93E-135	family with sequence similarity 46, member C
ENSECAG00000026268	0.715520798	37.32666	52.16712	5.70507	9.61E-19	2.96E-17	no information available
HBAZ-HORSE	7844.322539	409460.89798	52.19838	5.70593	5.81E-124	6.62E-121	Hemoglobin subunit zeta
TMC22	158.4243177	8813.03671	55.62932	5.79777	5.56E-146	1.45E-142	transmembrane and coiled-coil domain family 2
ENSECAG00000026903	0.933159143	53.86162	57.71965	5.85099	8.13E-20	2.74E-18	no information available
APOBEC3	0.238506933	14.02287	58.79440	5.87761	5.40E-08	6.90E-07	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 2
QZ1P9-HORSE	857.5927715	61294.93480	71.47324	6.15933	1.45E-168	8.82E-165	solute carrier family 4, anion exchanger, member 1
IGHE	0.370755057	27.23478	73.45762	6.19884	7.33E-15	1.71E-13	immunoglobulin heavy constant epsilon
FHDC1	9.861584866	753.87639	76.44576	6.25636	1.67E-113	1.60E-110	FH2 domain containing 1
ENSECAG00000018729	0.238506933	18.57966	77.89989	6.28355	2.28E-11	4.07E-10	no information available
SEC14L3	0.691205737	58.76199	85.01374	6.40962	1.60E-25	7.37E-24	SEC14-like 3
ACCN4	0.228072639	21.61158	94.75744	6.56617	2.01E-12	3.90E-11	amilonide-sensitive cation channel 4, pituitary
SUSD2	0.238506933	31.53612	132.22309	7.04683	3.12E-15	7.51E-14	sushi domain containing 2
ENSECAG00000025785	0.228072639	37.06228	162.50210	7.34431	2.81E-18	8.37E-17	no information available
ATP4A	0.922724849	408.66449	442.88879	8.79080	5.39E-85	2.65E-82	ATPase, H ⁺ /K ⁺ exchanging, alpha polypeptide
ENSECAG00000016466	0.228072639	188.04436	824.49327	9.68736	1.88E-54	3.56E-52	no information available

Table 7.6: The top twenty genes which were identified as down-regulated in the affected animals compared to the control group.

Pathway analysis using KEGG Mapper

Of the 1895 genes identified as displaying significant differential expression, 11 were involved in the haematopoietic cell lineage, B-cell receptor signalling or Jak-Stat pathway. Of these four were up-regulated and seven were down-regulated in the affected group with respect to the control group.





b)

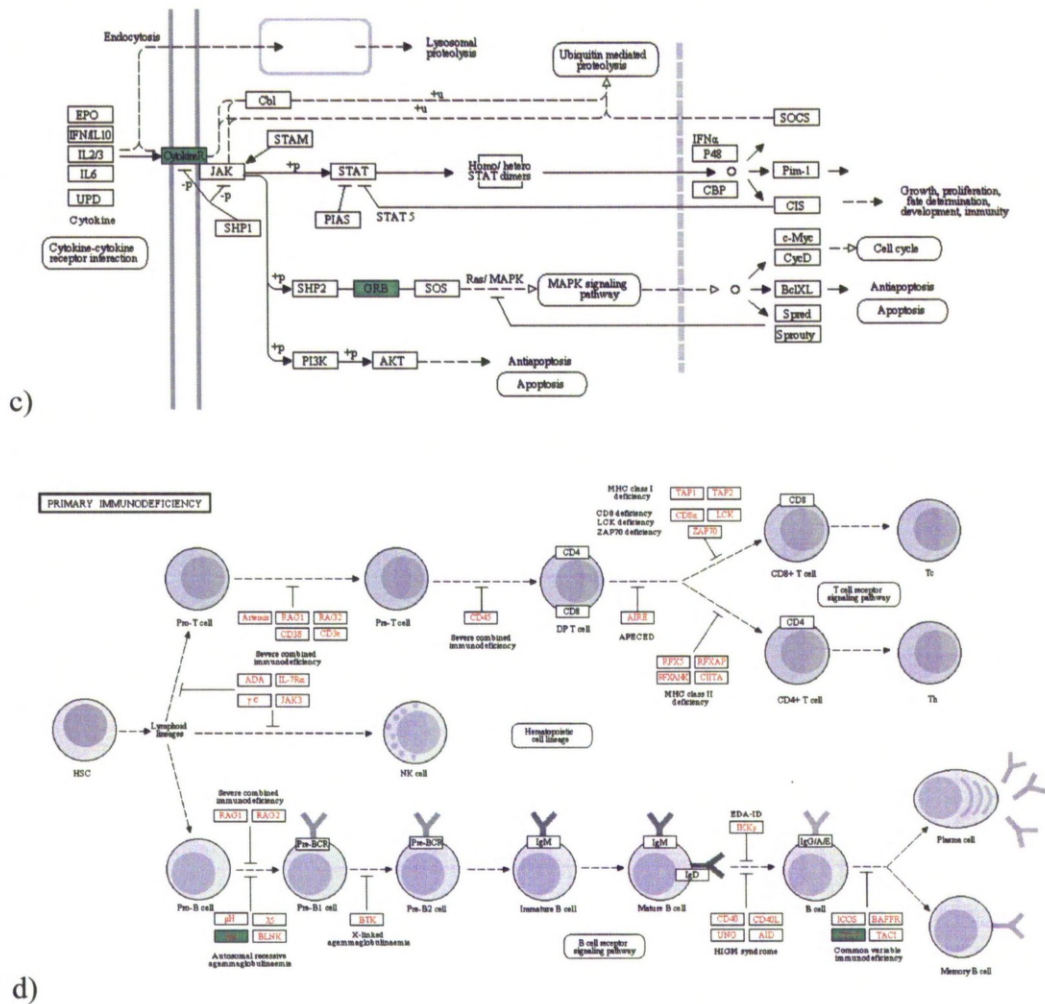


Figure 7.6: Pathway analysis using KEGG Mapper. Visual representation of three pathways which were deemed as those most likely to be affected by the FIS phenotype. Additionally the ‘primary immunodeficiency pathway’ is provided, which shows those genes which are implemented in other immunodeficiencies. Those genes shown in green are down-regulated in the affected group compared to the control group and those genes shown in red are up-regulated in the affected group compared to the control group. A) Haematopoietic cell lineage. B) B-cell receptor signalling. C) Jak-Stat pathway. D) Primary Immunodeficiencies.

Pathway analysis using Ingenuity

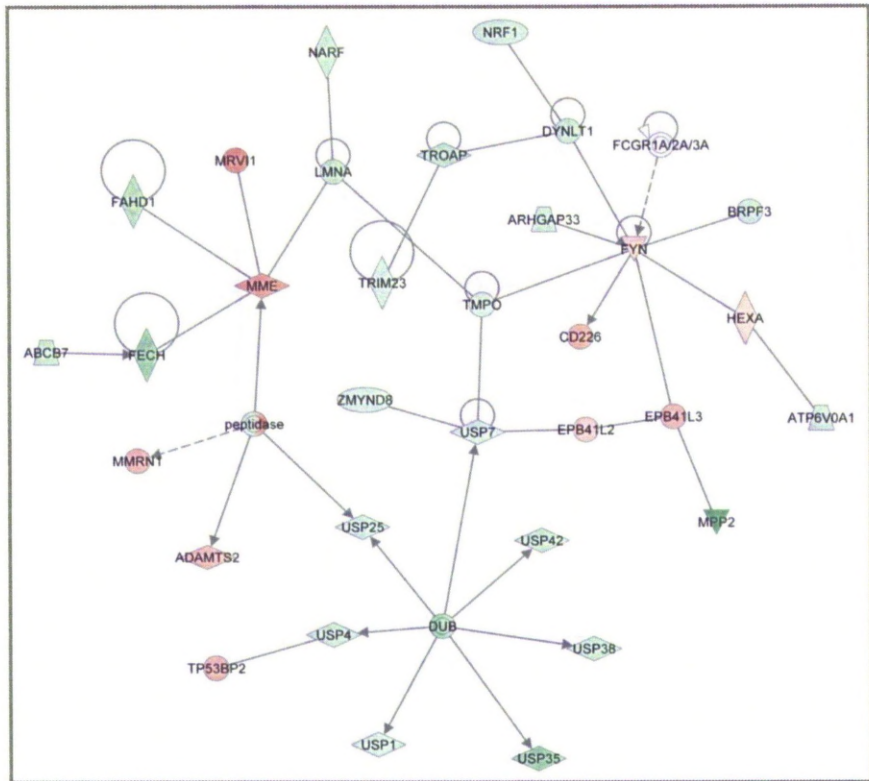
Those genes that were identified as differentially expressed, between the affected group and control group, were analysed using Ingenuity IPA software to identify gene relationships and pathways which were substantially affected by the disease

phenotype. Of the 1,895 differentially expressed genes, 1,270 were eligible for networking analysis and 1,215 were eligible for functional and pathway analysis. Those genes which were not eligible for analysis were those genes where literature on their function was not available. A total of 50 networks (Appendix 20) were created from the 1,270 proteins that were analysed, all with networking scores ≥ 12 (≤ 1 in 1000 chance of being due to random chance alone), with the three top networks achieving network scores of 39 (1 in 1×10^{39} chance of being due to random chance) (Table 7.4). There was substantial overlap of the genes within the 50 networks, resulting in an overrepresentation of a large number of functionally similar networks. Three of the top networks overlapped functionally, being involved in cellular assembly, the cell cycle and DNA repair and recombination, but showed little overlap in constituent genes. Network two (network score 39) contained 35 of the proteins significantly altered by FIS, with putative biological functions identified as ‘Post-Transcriptional Modification, DNA Replication, Recombination and Repair, and Molecular Transport’. This network contains 42 interactions, of which 95.2% are direct and 4.8% are indirect relationships. A visual representation of network two is shown in Fig 7.7. Network three (network score 39) contains 35 of the proteins significantly altered by FIS, with putative biological functions identified as ‘Cell Cycle, Embryonic Development and Cancer’. This network contains 51 interactions, of which 37.3% are direct and 62.7% are indirect relationships. A visual representation of network three is shown in Fig 7.7. Network four (network score 37) contained 35 of the proteins significantly altered by FIS, with putative biological functions identified as ‘Cellular Assembly and Organisation, DNA Replication, Recombination and Repair, and Cell Cycle’. This network contains 62 interactions, of which 96.8% are direct and 3.2% are indirect relationships. A visual representation of network four is shown in Fig 7.7.

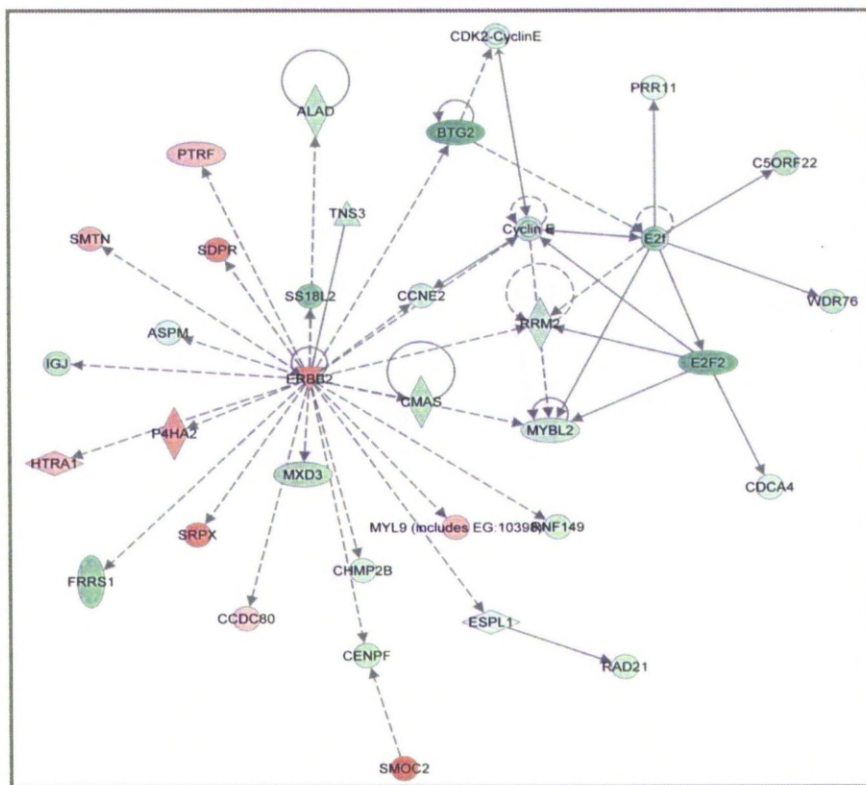
Top Networks

Network	Associated Network Functions	Network Score
1	Drug Metabolism, Amino Acid Metabolism and Small Molecular Biochemistry	39
2	Post-Transcriptional Modification, DNA Replication, Recombination and Repair, and Molecular Transport	39
3	Cell Cycle, Embryonic Development and Cancer	39
4	Cellular Assembly and Organisation, DNA Replication, Recombination and Repair, and Cell Cycle	37
5	Genetic Disorders, Ophthalmic Disease and Inflammatory Disease	33

Table 7.4: *The five associated networks which are most significantly altered by the FIS phenotype.*



a)



b)

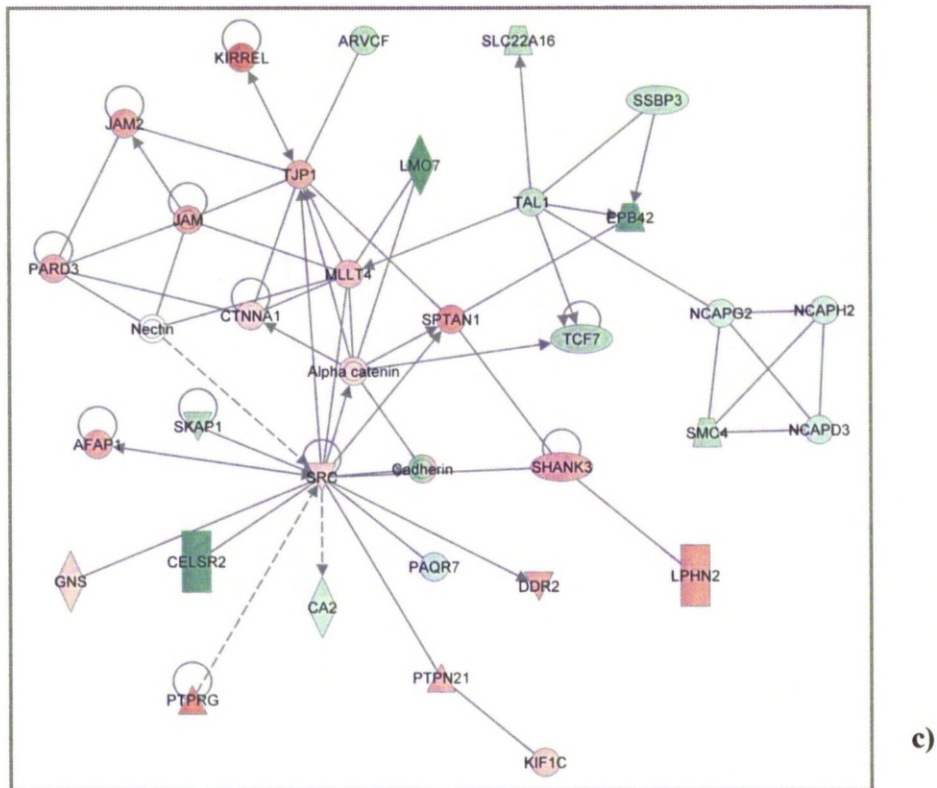
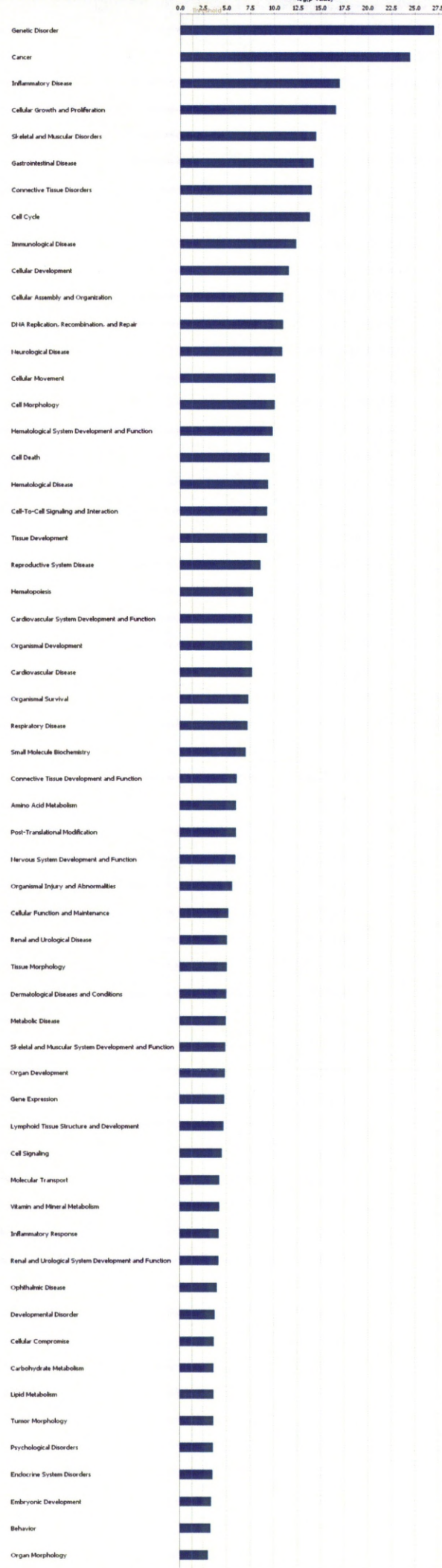


Figure 7.7: Visual representation of three protein networks generated by IPA, which are significantly affected by FIS. Each shape represents a protein and the adjoining line represents its association with another protein. Each shape represents a different type of molecule (horizontal diamond = peptidase, vertical diamond = enzyme, oval = transcription regulator, triangle = phosphatase, down-pointing triangle = kinase, pentagon = transporter and circles = other). Red molecules depict molecules that are up-regulated in FIS animals, green molecules represent those which are down-regulated in FIS animals and those in white are neither up or down-regulated. The intensity of the green and red molecules indicate the degree of down or up-regulation (a greater intensity represents a higher degree of up or down-regulation). **A)** Network two: Post-Transcriptional Modification, DNA Replication, Recombination and Repair, and Molecular Transport. **B)** Network three: Cell Cycle, Embryonic Development and Cancer. **C)** Network four: Cellular Assembly and Organisation, DNA Replication, Recombination and Repair, and Cell Cycle

P-values for biological functions were then generated by IPA, enabling the identification of those functions most significantly affected by the FIS phenotype. A total of 58 biological functions were identified as significantly (*P*-value ≤ 0.05) altered by FIS due to differential gene expression (Fig.7.8). Shown in Table 7.8 are the five most significantly altered biological functions within three functional group



headings ‘disease and disorders’, molecular and cellular functions’ and ‘physiological system development and functions’, all of which are closely related to the pathological presentation of an FIS foal.

Figure 7.8: *Biological functions and disease pathways which are significantly associated with the FIS phenotype, resulting from up and down-regulated protein-protein interactions.*

Diseases and Disorders

Name	P-Value	Number of genes
Genetic Disorder	$9.37 \times 10^{-28} - 1.06 \times 10^{-03}$	783
Cancer	$3.43 \times 10^{-25} - 1.06 \times 10^{-03}$	453
Inflammatory Disease	$1.08 \times 10^{-17} - 7.86 \times 10^{-04}$	396
Skeletal and Muscular Disorders	$3.59 \times 10^{-15} - 7.86 \times 10^{-04}$	393
Gastrointestinal Disease	$6.73 \times 10^{-15} - 4.32 \times 10^{-04}$	306

Molecular and Cellular Functions

Name	P-Value	Number of genes
Cellular Growth and Proliferation	$2.78 \times 10^{-17} - 1.06 \times 10^{-03}$	416
Cellular Development	$2.94 \times 10^{-12} - 1.06 \times 10^{-03}$	342
Cellular Organisation	$1.13 \times 10^{-11} - 9.71 \times 10^{-04}$	203
Cell Cycle	$1.63 \times 10^{-14} - 1.06 \times 10^{-03}$	194
DNA Replication, Recombination and Repair	$1.13 \times 10^{-11} - 5.89 \times 10^{-04}$	145

Physiological System Development and Function

Name	P-Value	Number of genes
Tissue Development	$5.38 \times 10^{-10} - 1.05 \times 10^{-03}$	223
Haematological System Development and Function	$1.41 \times 10^{-10} - 1.06 \times 10^{-03}$	187
Organismal Development	$2.05 \times 10^{-08} - 5.85 \times 10^{-04}$	176
Haematopoiesis	$1.76 \times 10^{-08} - 1.06 \times 10^{-03}$	119
Cardiovascular System Development and Function	$2.05 \times 10^{-08} - 5.85 \times 10^{-04}$	99

Table 7.8: *The most highly associated biological functions and disease pathways altered by the differential expression observed in FIS-affected foals when compared to healthy age matched controls.*

7.4 Discussion

The aim of this pilot study was to evaluate the experimental design, looking specifically at sampling techniques and extraction methods and to evaluate gene expression levels in FIS-affected foals compared to age matched controls. Additionally, networking analysis was performed to establish those biological functions and networks that were most disrupted in FIS foals. High quality results were obtained from this investigation, enabling a thorough evaluation of the study design, whilst yielding some valid conclusions about those networks and biological functions which are most disrupted in an FIS-affected foal.

Two different extraction protocols were used in this experiment; the first was used for extracting RNA from bone-marrow biopsies and the second was an RNA extraction method designed specifically for bone-marrow aspirates. Both extraction methods yielded high quality intact RNA; however, the PAXgene protocol was the more user-friendly and did not involve a manual homogenisation step like the other protocol. Although the PAXgene method was more user-friendly in the laboratory, these samples were haemodiluted, with a great deal of circulatory blood contamination, and histological examination revealed that only two of the seven samples contained bone-marrow tissue. Therefore, due to the variability of sample quality obtained from a bone-marrow aspirate, to ensure continuity between samples I would recommend bone-marrow biopsies as the sample of choice for future investigations.

Histological examination was performed on all of the bone-marrow aspirates, to confirm that they were representative of bone-marrow tissue. Of the seven samples, five were reported not to be representative of bone-marrow tissue, and although complete maturation sequences were not seen in the remaining two samples, there was some evidence of bone-marrow tissue so they were selected as control samples. A further control sample came from a bone-marrow biopsy which was taken post-mortem from a foal which was euthanized as it was suffering from colic. This sample was stored and processed according to the protocol which was used for the

affected foal samples. As part of the analysis, a heatmap was produced to provide a visual representation of the gene expression amongst the samples, indicating how similar or dissimilar the samples were. It would be expected that all of the FIS-affected samples would cluster together as would the control samples, showing a clear demarcation between the two groups. However, the unique sample which was a control sample taken from an FIS-unaffected foal post-mortem, clustered with the affected samples, showing that this sample was more closely related in terms of gene expression levels to the affected group than the control group. This suggests that because the animal was suffering from colic, a condition which would initiate an immunological response, that the bone marrow pathology of FIS animals is not too distinct from the pathology of colic, and indeed may not be all that distinct from the pathology of many diseases which cause an immune response. Additionally it may also suggest that either sampling technique or sample processing resulted in this animal clustering with the affected group and hence displaying abnormal gene expression. Based on this, it was decided that this sample would be excluded from pathway analysis.

High quality sequencing data was obtained for all seven samples, with an average of 74.49% of the reads mapping uniquely to the reference genome. The alignments were viewed in the UCSC genome browser, which enabled a comparison of the alignments of the samples. The alignments were specifically viewed in the FIS critical region to search for alternative splice variants and possible transcript truncation. There was an excellent match between the alignments of the affected animals and the controls, which provided further evidence that there are no mutations which lead to truncation of the genes within the FIS critical region.

Differential gene expression was assessed using transcript abundance levels. Prior to assessing differential gene expression between the two groups, the biological replicates within the two groups were compared to confirm that they had similar gene expression levels. This is because a core assumption of the analysis is that the mean is a good predictor of variance. At low counts where shot noise dominates, higher sequencing depth will improve the signal-to-noise ratio whereas for high counts, where the biological noise dominates, only additional biological replicates will improve the signal-to-noise ratio. This investigation used relatively few samples, particularly for the control group which comprised two samples, however

the estimation of variance for each sample revealed that the underlying assumption of DESeq, i.e. that the mean is a good predictor of variance, held true for this investigation. As a further assessment of gene expression levels across biological replicates, per-gene estimates of variance were calculated for each group. Plots of this data revealed that both groups followed the line of best fit extremely well, considering the small number of samples. The affected group had less noise than the control group, and this was probably due to higher sample numbers.

The two groups were then compared to identify those genes which were significantly differentially expressed. The affected-control comparison identified 1895 genes as significantly differentially expressed, of which 854 were up-regulated and 1041 were down-regulated. To identify those biased due to different sampling techniques and handling, two further comparisons were performed. The first compared the differential gene expression between the FIS-affected group and the unique sample; these had been processed and handled in the same way. However, due to the fact that the unique sample was from an animal which was suffering from colic and euthanized as was in a critical state, any conclusions drawn from this comparison have limited use. Although, this comparison revealed that the FIS-affected group and the unique group had similar gene expression levels, with only 187 differentially expressed genes. A second comparison compared the unique sample to the control group; these had been sampled using different methods and the unique sample was collected post-mortem while the controls were collected from live foals. Additionally, the control group were all healthy animals whereas the unique sample came from an animal which was suffering colic and in a critical state. This identified 233 differentially expressed genes - possibly fewer than expected and suggests that the sampling techniques may have had a minimal effect on observed gene expression variation. Based on these results, it does suggest that the 1895 differentially expressed genes identified between the FIS-affected and control comparison, are likely to be indicative of the gene expression differences which arise from FIS. Of these, 18 were not expressed at all in the control samples. Examination of the functional role of these 18 genes revealed that only two, *NRXNI* and *LBP*, had functional roles which could be directly associated with the FIS phenotype, suggesting that FIS has a large pathological impact on affected individuals, many of which changes are likely to be compensatory affects. Some foals appear to have a

peripheral ganglionopathy, a condition which affects the peripheral nerves. *NRXN1* is a member of the neurexin family, forming part of the receptor in the nervous system in vertebrates (Yue et al., 2011). The protein encoded for by *LBP* is involved in the acute-phase immunologic response to gram-negative bacterial infections and furthermore it plays an important role in the *LBP* dependent monocyte response. Monocytes are essential for the stress-response and immunological response, and the majority of monocytes are stored in the spleen (Lakatos et al., 2011), an organ which has an abnormal structure in FIS-affected foals. Although this investigation has successfully identified a number of genes as being significantly differentially expressed, identifying a direct relationship between the disease and the transcriptional disruption would be very challenging and due to the limited number of samples used in this study and the extreme sampling techniques that were used in this investigation, any conclusions from this investigation would be limited.

Finally, network and pathway analysis was performed to identify those gene pathways and cellular networks that were significantly affected by FIS. Initially KEGG Mapper was used to produce visual representation of those genes which were significantly disrupted in pathways which are most likely to be affected by FIS based on disease pathology. Of the 1895 significantly differentially expressed genes, 11 were identified as being involved in the haematopoietic, B-cell lineage and Jak-Stat pathways, four were up-regulated in the affected group and seven were down regulated. All of these genes besides two are involved in the immunological response and development of B-lymphocytes, with the two most significantly affected genes being *CD79A* (12-fold) and *CD79B* (20-fold), which were down-regulated in the affected group with respect to transcriptional levels in the control group. It would be expected based on the pathology of FIS and based on the results of the immunohistochemical staining using *CD79A* performed in Chapter 3, that expression of *CD79A* and *CD79B* would be significantly lower in affected animals compared to controls. This is because the antigen receptors on the surface of B-Lymphocytes contain a heterodimeric signaling component which is composed of *CD79A* and *CD79B*, with the presence of this receptor being essential for normal B-lymphocyte development and function (Kremyanskaya and Monroe, 2005). Therefore as FIS foals have severely depleted numbers of B-lymphocytes, we would expect to see much reduce expression of these two genes. A second feature of FIS is a progressive

anaemia which is unrelated to any blood loss or haemolysis. Two of the eleven genes were involved in the erythrocyte lineage of the haematopoiesis pathway, EPOR which is down-regulated by 3-fold in the affected samples, and CD36 which is up-regulated by 0.3-fold in the affected samples. EPOR encodes the erythropoietin receptor, with stimulation of this receptor being essential for the survival of erythroid cells. Mutations of EPOR that result in an up-regulation of this gene have been shown to be associated with familial erythrocytosis (over-production of erythrocytes) (Gombos et al., 2011). The second gene is CD36, which was up-regulated in the affected animals. Mutations in CD36 in humans has been highly associated with aplastic anaemia, a condition where the bone marrow does not produce sufficient cells numbers to replenish circulatory erythrocytes, lymphocytes and platelets (Trinh-Trang-Tan et al., 2010). Additional analysis using IPA, a computational based pathway analysis was used to identify those biological functions which are predicted to be significantly disrupted by the disease, providing further understanding of the clinical presentation. Of the 1895 differentially expressed genes identified, 1270 were eligible for networking analysis using IPA and 1215 were eligible for functional and pathway analysis. A total of 50 networks were identified as significantly altered by differential gene expression (network score 12 – 39), although there was a substantial overlap of genes within networks which resulted in overrepresentation of functionally similar networks. The top five networks identified by IPA were involved in metabolism, DNA replication, recombination and repair, the cell cycle, genetic disorders and inflammatory disease. Biological functions and gene pathways which were identified as most disrupted by FIS included the haematological system and its development, tissue development, cell growth and diseases including genetic disorders and inflammatory disorders. The main characteristics of FIS which have been revealed by clinical and pathological examination include a lack of thymic tissue, lack of germinal centres in secondary lymphoid tissues, peripheral ganglionopathy, tissue inflammation, anaemia and a severe B-lymphocyte deficiency. All of the clinical and pathological characteristics of FIS are associated with those gene pathways and biological functions which were identified as significantly affected by FIS, based on gene expression levels in FIS-affected foals. Therefore, this provides evidence that the causal variant has a knock-on effect, causing disruption to multiple gene pathways, which gives rise to the FIS phenotype. As discussed in chapter 5, a non-synonymous mutation in the single

exon gene *SLC5A3*, was identified as highly associated with FIS. This gene was shown to be up-regulated in FIS-affected foals, compared to healthy controls, by approximately five-fold. Networking and pathway analysis did not implicate this gene in any of the significantly disrupted pathways or networks, providing no further understanding of potential protein-protein interactions that could support this mutation as causal. However, IPA is limited to drawing conclusions and predicting protein-protein interactions based on current literature. There is little literature available on *SLC5A3* and therefore it was unlikely that this analysis would provide any novel predictions.

This experiment has highlighted how experimental design and the matching of cases and controls are essential to yield reliable results and valid conclusions. Although the bone-marrow aspirates were variable, the extraction method used for processing these samples was more user friendly in the laboratory and also obtaining the samples was a great deal easier than obtaining bone-marrow biopsies, as the procedure was less invasive. This is because the bone-marrow biopsies were obtained post-mortem from the femur of foals, which required sawing of the bone and scraping of the bone marrow from within the bone. Therefore this method is not conducive to live sampling from healthy foals as it would require the euthanasia of healthy animals. However, an alternative sampling method could be adopted for future studies, for example bone marrow puncture biopsies, which could be obtained from live animals under sedation with minimal effects to the animal. Further to this puncture biopsies would provide a more representative sample of bone marrow than bone marrow aspirates. Should bone marrow puncture biopsies not be a suitable sampling technique, then an improved experimental design would use bone-marrow aspirates from both controls and affected animals, obtaining sufficient numbers to allow for variability in samples. Also, a full cytological examination of samples would enable aspirates to be selected and matched based on their cellular composition.

Although every care was taken to ensure that affected samples were taken immediately post-mortem, it is likely that a degree of the differential expression will be a direct result of the pathological affects associated with death. Gene expression levels vary greatly between individuals, depending upon many internal and external factors. Therefore, by taking tissues for gene expression analysis at any one time,

only provides a snapshot of gene expression levels at that time point. For future investigations, serial sampling from the same animal may provide a more accurate picture of how gene expression is affected by this disease. Additionally, it would be suggested that samples be taken from live affected animals, which are not showing clinical signs of the disease. Affected animals could be identified at birth using the DNA-test described in chapter 6, and serial sampling performed over a set time-period. By doing so, any differential gene expression identified is more likely to be directly related to the disease. This would identify those pathways which are significantly disrupted, rather than providing a snapshot of the gene expression in a diseased animal which is undoubtedly suffering from secondary effects.

The results obtained from this experiment, considering how few samples were used, are of high quality and could undoubtedly be utilised in future investigations provided the additional samples are matched. Although further expression studies would enable a further understanding of the disease pathology, I feel a functional study, examining the effects of the *SLC5A3* mutation would be a more suitable natural progression at this time.

Chapter 8

General discussion and Future studies

	Page
Mapping the associated loci using genetic markers	216
Next-generation sequencing – A revolutionary tool for mutation mining	217
The sodium/glucose co-transporter family	218
Evaluation of the sodium/myo-inositol co-transporter as the causal mutation of FIS: Structural organisation and functional analysis	220
Breeding management to reduce carrier prevalence	227
Pathological and biological effects of FIS	229
Conclusion and final remarks	230

Chapter 8 – Final Discussion and Future Studies

This study has further characterised Foal Immunodeficiency Syndrome, and has led to the development of a DNA-based test which can be used to determine the genotype of individuals. The major novel findings from this study were:

- FIS has a genetic aetiology and pedigree analysis of FIS-affected Fell and Dales foals has led to the identification of a common ancestor to all of the affected foals.
- Identification of a highly associated mutation on chromosome 26, a non-synonymous SNP in the single exon gene *SLC5A3*, which segregates 100% with FIS. This mutation was expressed as a Mendelian trait (autosomal recessive), with a definitive phenotype in FIS-affected individuals.
- FIS affects both Dales and Fell Ponies, and has also transferred into the Coloured Pony population.
- FIS-carrier prevalence is approximately 38 - 48 per cent in the Fell Pony population and 10 – 18 per cent in the Dales Pony population.

Mapping the associated loci using genetic markers

Microsatellite markers have long been used to successfully map Mendelian traits in the horse (Swinburne et al., 2002, Terry et al., 2004, Tryon et al., 2007) and therefore linkage analysis using microsatellite markers was adopted here to map FIS. This led to the identification of a single locus which showed significant association (LOD score >3) to FIS. Additionally, homozygosity mapping was performed, to identify markers which displayed a loss of heterozygosity compared to the control group. This also identified the same marker, on chromosome 26 as displaying a significant loss of heterozygosity in the affected group. Due to limited marker numbers on some of the chromosomes, it was felt that in order to definitively confirm the location of the FIS mutation, further genetic investigation was required.

The release of the equine genome in January 2008 facilitated the development of an equine SNP Beadchip, which enabled the interrogation of ~54,000 genetic markers, for genetic association studies. SNP-based whole genome studies have proved to be highly successful in mapping simple Mendelian traits in other domesticated species (Karlsson et al., 2007, Charlier et al., 2008). Therefore, the release of the EquineSNP50 Beadchip provided the equine community with a powerful new tool, providing an alternative approach for disease mapping. Initially, it was anticipated that an equine SNP Beadchip of ~100,000 SNPs was required to enable the successful mapping of disease loci across all breeds (Wade et al., 2009). However, here we demonstrated how the EquineSNP50 Beadchip could be used to perform a successful genome-wide association study, which led to the successful mapping of the FIS locus. Although ~10,000 SNPs were excluded due to a minor allele frequency of less than 2%, a single, unambiguous, disease-associated signal which supported the results from the microsatellite linkage mapping study was obtained, clearly identifying chromosome 26 as the location of the FIS mutation.

The EquineSNP50 Beadarray has provided an excellent alternative tool for mapping traits in the horse. However, for complex studies which investigate polygenic traits, much larger sample sets and an increased marker density would be required. In March 2011 a new, markedly improved EquineSNP74 Beadchip was released (Mickelson, 2011), which may now provide sufficient density and power to identify loci which are linked to complex traits.

Next-generation sequencing – A revolutionary tool for mutation mining

The target region on chromosome 26 was captured using NimbleGen sequence capture arrays, an alternative method to the more traditional based PCR methods. Since this investigation was performed, solution based methods have been developed for selection of target sequences at the bench. These methods offer a more cost effective approach to selection of the target sequence. Although sequence capture based methods offer an alternative, quicker method, they are not without limitations. They are open to selection and amplification bias, which could result in potential variants being missed. Therefore, careful consideration should be given to the selection of the method used for targeted selection of the critical region.

Next generation sequencing is a revolutionary tool, which enables the interrogation of entire critical regions when mining for disease mutations. This circumvents the more traditional approach of individually sequencing candidate genes using capillary based sequencing. The Roche 454 titanium series was used to re-sequence the FIS critical region, to identify a genetic mutation which segregated with the disease phenotype. This investigation was the first to utilise high-throughput re-sequencing to successfully identify a mutation which is highly associated with a genetic disorder in the horse. The variant was a single non-synonymous SNP which resulted in a proline to leucine substitution in the single exon of *SLC5A3*. No other variants that segregated with the disease were identified.

Single-end reads were obtained from this experiment, as paired-end sequencing was not available at the time. Identification of large scale rearrangements can be extremely difficult with single-end reads, due to the fact that any repeats will simply stack on top of one another during the alignment. Although every effort was made to exclude the possibility of large scale rearrangements in the FIS-critical region, it is possible that they could have been overlooked by the analysis pipeline. Therefore, further studies to assess large scale rearrangement would be required to exclude this as a possibility. Additionally, due to the NimbleGen probe design, small regions of the target region were not captured. Traditional Sanger sequencing methods were used to sequence these regions, although a small number still remain uncovered. To definitively exclude these regions, further studies should be conducted, to provide complete coverage of the FIS critical region on chromosome 26.

The sodium/glucose co-transporter family

The sodium/glucose cotransporter family (*SLC5A*), sometimes referred to as the sodium/substrate symporter family, has more than 220 members, of which 12 are human genes. Although gene structure is diverse between *SLC5A* family members, homology amongst the proteins is very high, with all but *SLC5A5* and *SLC5A11* having 14 transmembrane helices (Wright and Turk, 2004). The functions of all 12 genes have been previously described, based on expression studies, with genes being expressed in a wide range of unexpected tissues and cells. These functional studies have identified several of these genes as having surprising properties, with functional

properties differing greatly between species. Of the 12 members, five are plasma membrane sodium substrate complexes for the transport of glucose; one is a sodium substrate complex for the transport of lactate; one is a glucose activated ion channel; one is a sodium substrate complex for the transport of myo-inositol; one is a sodium substrate complex for the transport of iodide; one is a sodium/chlorine/choline cotransporter; one is an anion transporter and another is a sodium substrate complex for the transport of biotin, lipoate and pantothenate.

Although the precise functional relevance of these genes is not yet fully understood, *SLC5A8* has been implicated as a tumour suppressor gene in colorectal cancer (Gupta et al., 2006) and three further members of the SLC5A family have been identified as causing genetic diseases in man:

- Glucose-galactose malabsorption, arising from autosomal recessive mutations in the *SLC5A1* gene, presents as chronic diarrhoea in newborn infants (Turk et al., 1991). Mutations in this gene are associated with defective sugar transport. Missense mutations prevent critical conformational changes in the protein which cause trafficking defects.
- Autosomal recessive mutations in *SLC5A5* have been identified as the cause of congenital iodide transport defect, which results in defective thyroid hormone production (Pohlenz and Refetoff, 1999). Deletions and non-synonymous mutations in this gene have been identified, resulting in either a non-functional protein, a change in the 3-dimensional shape of the protein which impairs the function of the protein, preventing it from being positioned in the membrane, resulting in the impairment of iodide transport. Consequently, affected individuals often present with an enlarged thyroid which is attempting to compensate for the lack of hormones.
- The most recent *SLC5A* genetic mutation to be described is in the *SLC5A2* gene, and results in a condition known as familial renal glucosuria (Santer et al., 2000). Inherited as a dominant trait, this condition is characterised by persistent glucosuria, resulting from defective glucose absorption in the kidneys.

Evaluation of the sodium/myo-inositol co-transporter as the causal mutation of FIS: Structural organisation and functional analysis

Through the use of cloning, sequencing, mRNA analysis, and reporter gene assays, the genomic structure, transcription start site, polyadenylation signals, and promoter of the human *SLC5A3* has been confirmed. In the human, *SLC5A3* consists of a single promoter and two exons, spanning a region of approximately 26 Kb. Exon 1 contains 175 bp of 5' untranslated sequence and is 15 kb upstream of exon 2. The 9.5-kb exon 2 contains the entire 2157-bp open reading frame and a large 3' untranslated sequence with seven putative polyadenylation signals (Mallee et al., 1997). Located on human chromosome 21, *SLC5A3* shares the same genomic region as *MRPS6*, with the first exon of *SLC5A3* being shared also by *MRPS6* (Buccafusca et al., 2008) and the second exon of *SLC5A3* being embedded within the first intron of *MRPS6*. The first exon of *MRPS6* contains part of the coding region of the *MRPS6* gene, but not of the *SLC5A3* gene. The reading frame of the sodium/inositol cotransporter is embedded within intron 1 of the *MRPS6* gene (Fig: 8.1). In the equine genome, *SLC5A3* is predicted to consist of a single exon approximately 26 Kb upstream of the gene *MRPS6*, therefore in contrast to the human structural organisation of these genes, *SLC5A3* is not embedded within the *MRPS6* gene and *SLC5A3* consists of only one single exon.

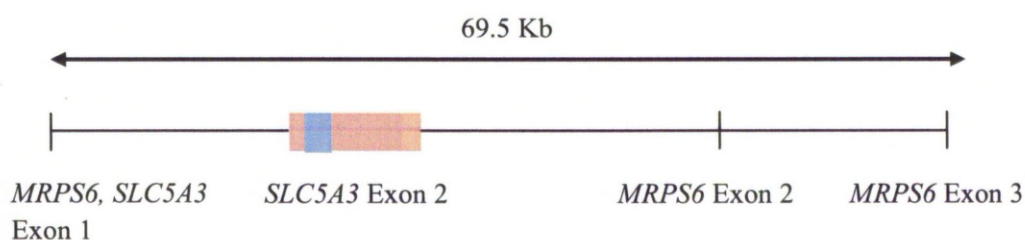


Figure 8.1: Structural organisation of *SLC5A3* and *MRPS6*. Both genes share the same genomic region on human chromosome 21, sharing the same exon 1.

The sodium/myo-inositol co-transporter (*SLC5A3*) was identified in this investigation as highly associated with the Foal Immunodeficiency Syndrome phenotype. The identified mutation is a non-synonymous SNP which segregates

100% with this disease. *SLC5A3* has been identified as responsible for maintaining a cellular concentration of the osmolyte myo-inositol (MI). In addition, *SLC5A3* also plays a crucial role in osmoregulation of cells, enabling the cell to maintain cell volume (Mallee et al., 1997). Equine *SLC5A3* is located on chromosome 26, with its single exon spanning approximately 1.6 Kb (30,658,870-30,660,513), with a protein of 541 amino acid residues. The FIS-associated SNP causes a proline to leucine amino acid substitution at residue 446 (equivalent residue 451 in the human protein), a residue which is conserved in all 12 eutherian mammals for which high-coverage sequence is available (sequence alignments shown in Fig: 8.2). Homology of the protein sequence is very high, with alignments of the *SLC5A* family members revealing that this proline residue is conserved in 11 of the 12 equine and human *SLC5A* genes, indicating functional significance.

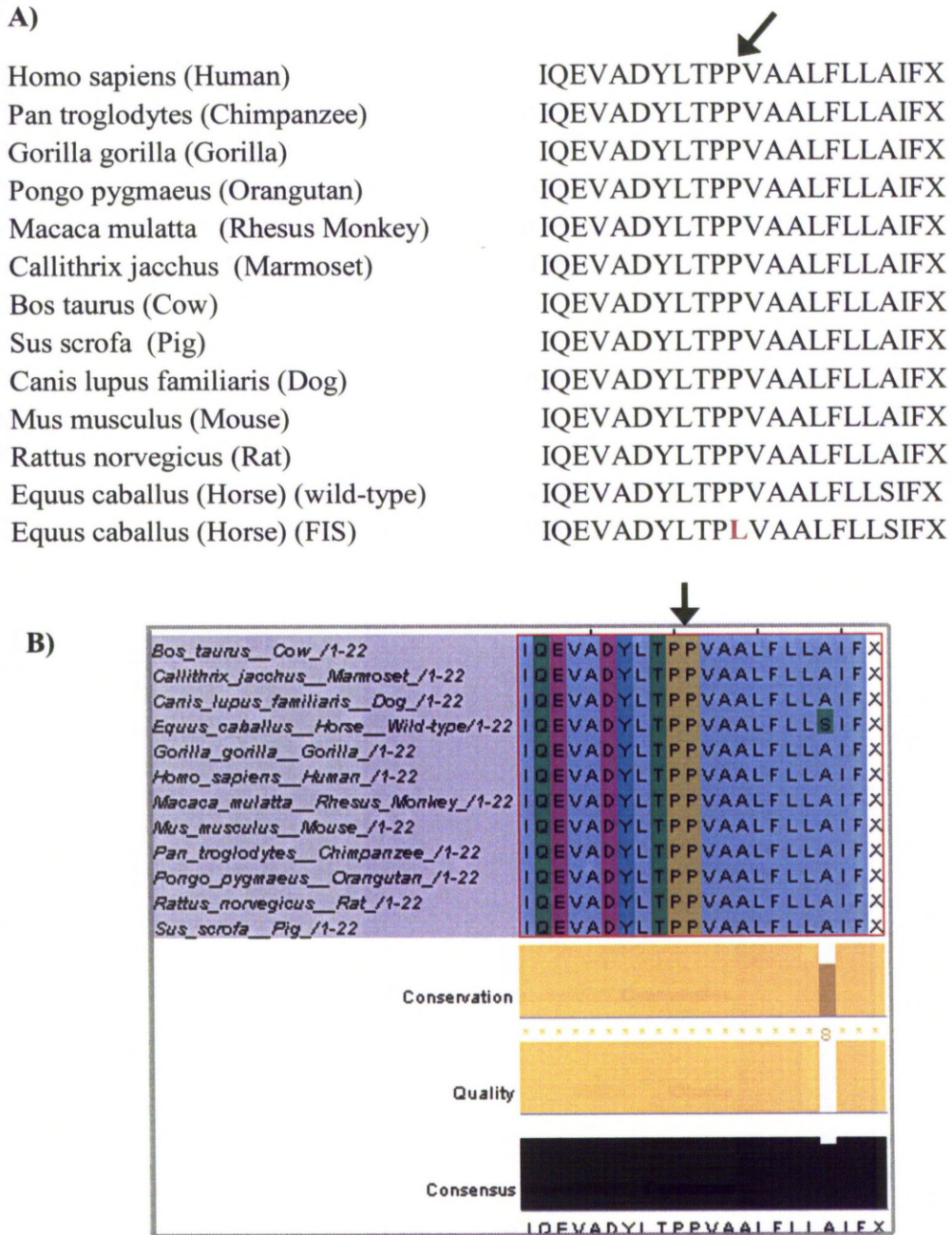


Figure 8.2: Amino acid alignment of *SLC5A3* in the 12 eutherian mammals for which high-coverage sequence is available. A) The proline residue (indicated by the arrow in both images), at position 451 in the human (equivalent residue 446 in the equine protein) is substituted by a leucine residue in FIS-affected animals. B) Amino acid alignments performed using Tcoffee online sequence alignment tool (http://www.ebi.ac.uk/Tools/services/web_tcoffee/toolform.ebi) show high conservation across species, with the proline residue indicated by the arrow, at position 451 in the human (equivalent residue 446 in the equine protein) being conserved in all 12 species.

The secondary structure of *SLC5A1* has recently been described in *Vibrio parahaemolyticus*, providing information on the structural changes which are adopted for the transport of small molecular weight solutes across the membrane (Faham et al., 2008). Alignment of the protein sequences of *SLC5A1* and *SLC5A3* suggests that the P446L substitution in equine *SLC5A3* is located in the eleventh transmembrane helix (Fig 8.3), which is involved in forming the substrate binding cavity (Faham et al., 2008). Membrane buried proline residues are commonly observed in transport proteins, suggesting that they have functional significance for the transport of molecules across the membrane (Brandl and Deber, 1986). Proline residues acts as hinges, bringing about crucial conformational changes, with induced proline mutations being shown to have a detrimental effect on the function and folding of proteins (Hiniker et al., 2006). Further, proline induced mutations in the transmembrane helix have been shown to have a profound effect on the function of an enzyme, reducing substrate affinity or diminishing substrate transport (Vilsen et al., 1989).

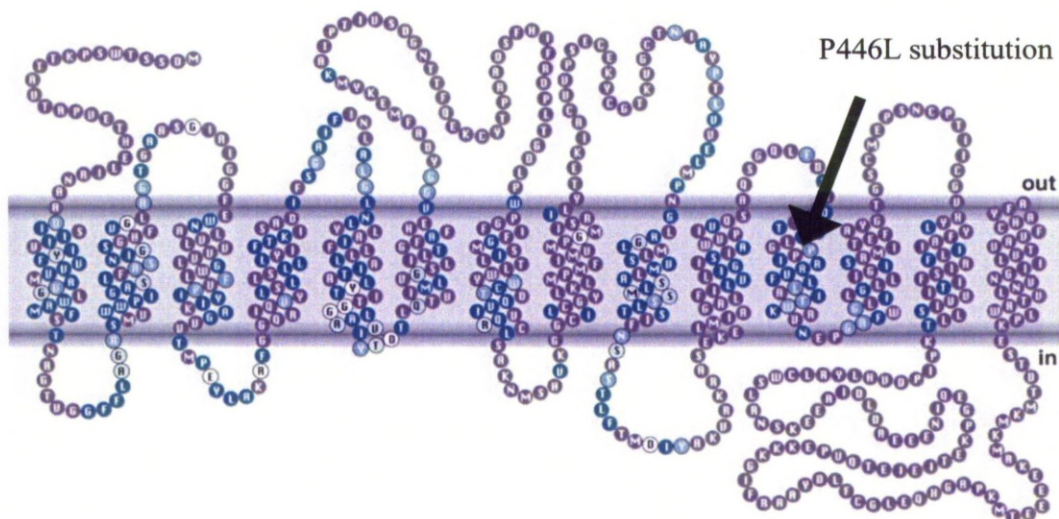


Figure 8.3: Secondary structure model for the human *SLC5A1* gene, which has 14 transmembrane helices. Alignment of the human *SLC5A1* and equine *SLC5A3* protein sequences predicts that the proline to leucine substitution, observed in FIS-affected individuals, falls within the eleventh transmembrane helix (the P446L substitution affects the proline residue shown in purple as indicated by the arrow). Image reproduced from Wright and Turk, 2004.

Although the precise importance of *SLC5A3* remains unknown, it has been shown that its primary responsibility is to maintain isotonicity by importing myo-inositol (MI) into cells and tissues. MI is transported into cells using the electrochemical gradient of sodium across the plasma membrane, and is an essential component of living cells. Further, MI is the precursor of phosphatidylinositol, whose derivatives are important for many normal biological functions, including cell survival, growth, vesicular trafficking and glucose homeostasis (Chau et al., 2005). Functional studies of *SLC5A3* have revealed that regulation of MI concentration during embryonic and foetal life is essential for the normal development of peripheral nerves. Mice which are homozygous null for *SLC5A3* die shortly after birth as a result of respiratory failure, due to the malformation of the nerves which are responsible for controlling breathing (Chau et al., 2005). One explanation for this is that the formation of phosphatidylinositol is inhibited due to the lack of MI, its precursor, which results in signalling abnormalities which are essential for the normal development of the nervous system (Berry et al., 2003). In addition to the defects in the peripheral nerves, *SLC5A3* null mice are also smaller and the curvature of the vertebral column is smaller (Chau et al., 2005).

SLC5A3 is an osmotic stress response gene, maintaining osmotic pressure and protecting cells from hypertonic stress, through the accumulation of the osmolyte myo-inositol in the cell. Maintaining the volume of cells by the transfer of osmolytes is crucial for preventing dehydration of cells which results from an increased osmotic pressure in the extracellular environment. Prevention of osmolyte accumulation has been shown to result in cellular dehydration which consequently disrupts normal cell function (Kwon et al., 1992), resulting in the denaturation of intracellular molecules and damage to sub-cellular architecture (Haussinger, 1996). Mammalian cells are not normally subjected to extreme hypertonic stress because complex and sensitive systems have evolved to maintain and control homeostasis. However there are two exceptions to this, the first is the kidney, which is exposed to extreme hypertonic stress as part of the urine-concentrating mechanism, and the second is the lymphoid tissue microenvironment, where controlling osmotic stress has been shown to be important for the development of lymphocytes (Go et al., 2004).

The mechanism by which intracellular signalling mediates the osmotic stress response in mammalian cells has not yet been completely defined. However, it is

thought to involve a signalling cascade which leads to the transcription of those genes responsible for the accumulation of osmolytes. When the cell is exposed to osmotic stress, Rho-type guanine nucleotide is stimulated, initiating a signalling cascade which attracts cJun Kinase-interacting protein 4. The cJun Kinase-interacting protein 4 attaches to the Rho-type guanine nucleotide which in turn stimulates p38 Mitogen-activated protein kinase (p38 MAPK). Finally, the tonicity-response enhancer, Nuclear Factor of Activated T-cells 5 (NFAT5) is stimulated and binds to the complementary binding sites in the promoters of the osmolyte accumulation genes (Kino et al., 2010), of which *SLC5A3* is one (Fig 8.4). Binding of NFAT5 to the promoter activates transcription of these genes, so that they transport osmolytes such as myo-inositol into the cell, to maintain isotonicity of the cell, protecting it from dehydration (Burg et al., 2007).

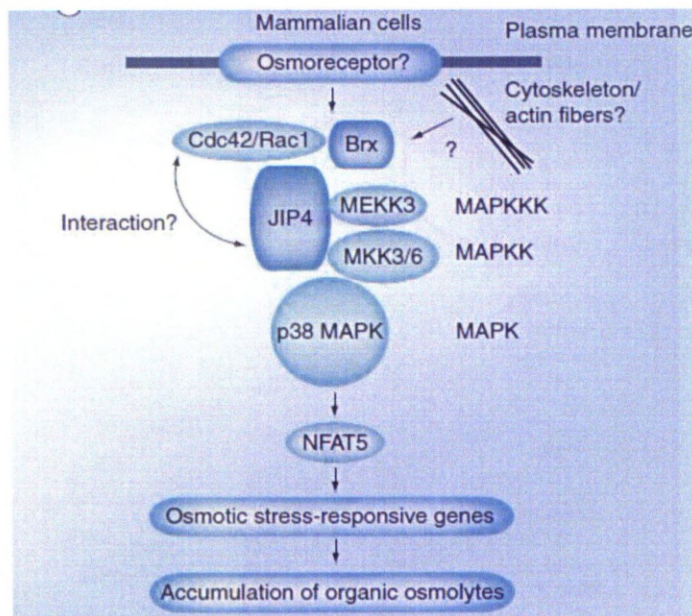


Figure 8.4: Proposed signalling cascade for the activation of the osmotic stress response genes, which stimulates the accumulation of osmolytes in the cell to protect the cell from dehydration. (Kino et al., 2010)

Although the precise biological relevance of the osmotic stress response in immune tissues is unknown, it is widely accepted that lymphoid tissues are hyperosmotic relative to blood and other tissues in the body (Go et al., 2004). NFAT5 is the only known osmoprotective transcription factor and is known to be highly expressed in

the thymus and induced upon lymphocyte activation, with recent studies indicating that the osmotic microenvironment of the thymus and spleen is critical for the normal development of lymphocytes (Go et al., 2004). Knockout studies in mice have shown that complete loss of NFAT5 results in late gestational lethality, whereas those with partial loss of NFAT5 function were viable with defective adaptive immunity and a substantially reduced spleen and thymus (Go et al., 2004). Thus, this observation shows that the normal osmotic stress response pathway is crucial for the normal development of immune tissues and lymphocytes.

Specific cell types have precisely defined acute levels of hypertonicity that they can tolerate. Failure to maintain osmotic pressure of cells will quickly lead to increased levels of NaCl, which will subsequently denature the cell, resulting in a loss of function or apoptosis (Burg et al., 2007). Cells are particularly susceptible to osmotic stress when they are undergoing cell proliferation, a process which lymphocytes rapidly undergo. Exposure of cells to osmotic stress results in perturbing effects, including cell cycle arrest, DNA damage and apoptosis, all of which render the cell dysfunctional. Further, mature B-lymphocytes and macrophages have been shown to be highly susceptible to osmotic stress, resulting in apoptosis, which suggests that these cells types produce their own death ligand (Burg et al., 2007).

A primary feature of FIS-affected foals is a profound B-lymphocyte deficiency and anaemia, with severe depletion of lymphoid tissues, including the thymus and a lack of germinal centres in the spleen. Although the precise relevance of osmotic stress in the lymphoid microenvironment is unknown, it has been shown that impairment of the signalling cascade required for maintaining isotonicity results in defective adaptive immunity and substantially reduced spleen and thymus (Go et al., 2004). Further to this, B-lymphocytes, which are severely depleted in FIS-affected foals, are known to be highly susceptible to osmotic stress, resulting in apoptosis of these cells (Burg et al., 2007). A secondary feature of FIS foals is a peripheral ganglionopathy, which has been observed in some, but not all FIS-affected foals (Scholes et al., 1998). Although the precise biological relevance of *SLC5A3* has not been described, it has been shown that complete knockout of this gene results in the death of mice shortly after birth due to hypoventilation (Berry et al., 2003), which is most likely

due to the failure of the peripheral nervous system (Chau et al., 2005). There has been no literature published explaining the effects of osmotic stress on the development of erythroid precursors and therefore linking the profound anemia suffered by FIS-affected foals to the osmotic stress pathway and the mutation identified in *SLC5A3*, is more complex.

The mutation observed in FIS-affected foals is a proline substitution. Proline residues have been shown to have functional significance in transmembrane helices, bringing about the conformational changes required for the transport of molecules across the membrane (Brandl and Deber, 1986). It is therefore proposed that the P446L substitution in *SLC5A3* results in a subtle change which impairs the substrate-binding complex of the transmembrane helices. Consequently, cells are unable to adequately regulate osmotic pressure leading to apoptosis; this particularly affects rapidly proliferating cells. Further functional studies are now required to demonstrate the differences between the Pro446 and Leu446 forms of the protein. This could be achieved by introducing this point mutation into transgenic mice and assessing transport function and the development of these mice. This study would enable further investigation into the biological effects of the mutation, ultimately providing an understanding of how the mutation results in a profound B-lymphocyte deficiency and anemia. In addition, an investigation into the effect of osmotic stress on haemopoietic stem cells, the progenitor cell type to both lymphocytes and erythrocytes, would also help to further explain how the *SLC5A3* mutation results in the FIS phenotype.

Breeding management to reduce carrier prevalence

Genetic screening tests have been developed for several inherited diseases in the horse (Tryon et al., 2007, Brooks et al., 2010), and with careful breeding management, have enabled the dramatic reduction of the carrier prevalence within the population. The identification of the novel *SLC5A3* mutation in FIS has been verified as segregating with carrier and affected horses and was therefore deemed suitable as the basis for a genetic screening test. Use of this test could ultimately lead to the eradication of this lethal mutation from the equine population. However, this

will require careful management over an extended time period, particularly in the Fell Pony breed, where carrier prevalence is estimated to be as high as 50%.

Breeding management in the Fell Pony should avoid the exclusion of carriers from breeding programmes as this would restrict the gene-pool and could lead to further genetic problems. Therefore, while carrier-carrier mating should be avoided to prevent the loss of foals from FIS, carrier-clear mating should be encouraged to enable the continued use of carriers, until such lines are replaced by clear animals. Carriers should be removed very slowly from the breeding population; this could be accomplished by replacing older carrier mares and stallions, with clear offspring that are of a comparable quality. Over time, this method would see the gradual decline in the carrier prevalence, leading to a point where all remaining carriers could be excluded.

Both the Dales and Coloured pony population have relatively low carrier prevalence compared to the Fell Pony. Therefore, all carriers in the Coloured pony population should be excluded from breeding and all interbreeding with Fell and Dales should only be performed with confirmed FIS-clear individuals. Pedigree analysis of the Dales population revealed that FIS-carriers are more prevalent in some lines. To avoid the loss of these lines, carrier-clear matings should be encouraged until older carrier animals have been replaced with clear animals of a similar quality. However, in the majority of situations, clear-clear matings should be encouraged.

Further studies are now required to provide an estimate of the FIS-carrier prevalence in the Coloured pony. This is because the current study only screened a relatively small population, which was a mixture of traditional gypsy cobs, which are known to have interbred with the Fell and Dales, and those ponies which are registered with the Coloured Horse and Pony Society. It is therefore likely that the Coloured pony population that was tested represents two different pony types, of which one may be at a greater risk of the FIS-mutation. The traditional gypsy cob is known to have regularly interbred with the Fell and Dales, as many are turned-out onto the Cumbrian Fells, to roam freely, and then collected at a later date, often with foals at foot (T. Capstick, personal communication). Whereas, ponies registered by the Coloured Horse and Pony Society represent a type which is more commonly seen in the show ring. Such a study would not only provide further details on the FIS-carrier

prevalence but also provide further information on those pony types which are at greater risk of carrying the FIS-mutation. Additionally, further press on this disease would help to make breeders more aware, alerting them to the fact that this disease is not limited to the Fell Pony, as was once thought. Finally, additional screening of Fell and Dales ponies should be considered to estimate FIS-carrier prevalence in populations other than those in the United Kingdom.

Pathological and biological effects of FIS

A pilot study was performed to investigate the global transcription of FIS-affected foals compared to healthy controls. Ultimately a full experiment would provide further understanding into the biological pathways which are disrupted by FIS and how this relates to the clinical characteristics of the disease. This investigation highlighted those areas which would need further refining in a full experiment but also provided some interesting results. In this experiment, bone marrow samples were used and therefore it should be acknowledged that the data from this investigation provides a 'snapshot' of the transcriptional level in these individuals at the time of sampling across all cell types in the bone marrow. Additionally, all analysis was performed on 'average' transcriptional levels across the two group, rather than individuals themselves. Pathways that were identified as most significantly disrupted included the haematological system and its development, tissue development and cell growth. Clinical characteristics of FIS include a severe B-lymphocyte deficiency (Thomas et al., 2005), progressive profound anaemia, peripheral ganglionopathy, absent germinal centres in the lymphoid tissues and severely depleted thymus (Scholes et al., 1998). The data obtained from this investigation does show that FIS-affected foals have significant transcriptional disruption to many pathways, including those pathways which are most closely associated with haematopoiesis and B-cell development, thus supporting the pathological findings of this disease.

Genetic mapping performed during this study revealed a single mutation in *SLC5A3* as highly associated with the FIS phenotype. However, it should not necessarily be expected that the gene which harbours the causal variant will show differential expression. This experiment identified a five-fold increase in the expression of

SLC5A3 in FIS-affected foals with respect to healthy age-matched controls. This suggests that there may be a feedback response controlling *SLC5A3* transcription which is attempting to compensate for the reduced transport of myo-inositol (Burg et al., 2007), probably due to the mutation leading to defective *SLC5A3* function. Pathways which were identified as significantly disrupted in FIS-affected foals included tissue development and cell growth. Both tissue development and cell growth, particularly with respect to the development of the lymphoid tissues and lymphocytes have been shown to be significantly affected by FIS. It can therefore be hypothesised that these pathways are significantly affected by the described *SLC5A3* mutation. Further investigations are now required to provide an understanding of the biological importance of *SLC5A3* and the networks that this gene interacts with in the context of FIS. Further transcriptional studies should therefore not be considered until gene specific functional studies have been completed.

Conclusion and final remarks

This work has provided important new information concerning the genetics of Foal Immunodeficiency Syndrome. FIS is a novel disease whose equivalent has not been described in any other species. Further, the mutation identified as highly associated with FIS is of biological interest because this research provides a novel hypothesis into the functional relevance of this gene. Functional studies of this gene may provide new information on its interaction with erythroid and lymphoid cell lineages. In turn, this may provide a new insight into foetal development, the immune system and the haemopoietic system.

Ultimately, the most successful and positive outcome of this investigation is the development of a genetic test which can be used to identify FIS-carriers and confirm the diagnosis of FIS-affected foals. As a result, the suffering of foals from this lethal disease can immediately be halted, and in time this mutation eradicated from the equine population.

Appendices

	Page
Appendix 1	
Sample summary sheet	231
Appendix 2	
DNA extraction protocol for whole blood samples	233
Appendix 3	
Whole genome microsatellite marker panel	235
Appendix 4	
Microsatellite markers used for linkage analysis and homozygosity mapping	237
Appendix 5	
Soft tissue extraction protocol	241
Appendix 6	
SNPs used for fine-mapping the FIS-critical region	242
Appendix 7	
Heterozygous high quality differences detected in sample 03_08 and the genotyping results for these SNPs in 36 additional samples	243
Appendix 8	
Cytology report for the bone marrow aspirates	245
Appendix 9	
UCSC alignments from the RNA-seq experiment – alignment of the FIS-critical region	247
Appendix 10	
Top 50 networks which were identified as significantly altered in FIS-affected animals from the RNA-seq experiment	250
Appendix 11	
Primer details: Primers designed specifically for this investigation	253
Appendix 12	
Abbreviations, buffers and reagents	261

APPENDIX 1

SAMPLE SUMMARY SHEET

Sample Name	Breed	Phenotype	Collected by:	Extracted by:	Sample type:	Microsatellite mapping	GWAS	Fine mapping (including candidate gene interrogation)	454 Re-sequencing (including interrogation of variants)	Global Transcriptionics
19 03	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x	x			
138 00	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
157 00	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
154 00	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
28 00	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
155 00	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
86 02	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x	x			
66 02	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x	x			
84 00	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x	x			
12 08	Fall	Adult Unknown	Author	Author	Blood		x			
479	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
154 00	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
157 00	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
159 00	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
69 66	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
70 69	Fall	Adult Unknown	G.Thomas	Author	Blood		x			
05 09	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
02 00	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
108	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x*				
26 02	Fall	Adult Unknown	G.Thomas	G.Hill	Blood					
21 01	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x				
19 03	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
79 00	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x				
472	Fall	Adult Unknown	G.Thomas	G.Hill	Blood				x	
108 00	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x				
05 03	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x				
28 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
27 09	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
28 09	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
52 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
51 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
54 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
55 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
59 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
55 02	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
44 03	Fall	Adult Unknown	G.Thomas	G.Hill	Blood	x				
25 01	Fall	Adult Unknown	G.Thomas	Author	Blood				x	
157 00	Fall	Adult Unknown	G.Thomas	Author	Bone Marrow		x			
20 09	Fall	Foal Unknown	Author	Author	Bone Marrow					x
21 09	Fall	Foal Unknown	Author	Author	Bone Marrow					x
04 08	Fall	Foal Unknown	Author	Author	Tissue Bone Marrow				x	
01 08	Fall	Affected	Author	Author	Tissue		x			
02 08	Fall	Affected	Author	Author	Tissue		x			
03 08	Fall	Affected	Author	Author	Tissue		x			
05 08	Fall	Affected	Author	Author	Tissue		x			
06 08	Dales	Affected	Author	Author	Tissue					
09 08	Fall	Affected	Author	Author	Tissue					
97 00	Fall	Affected	G.Thomas	G.Hill	Blood	x				
43 02	Fall	Affected	G.Thomas	G.Hill	Blood	x				
39 98	Fall	Affected	G.Thomas	G.Hill	Blood	x				
32 98	Fall	Affected	G.Thomas	G.Hill	Tissue	x				
32 98	Fall	Affected	G.Thomas	G.Hill	Blood	x				
16 01	Fall	Affected	G.Thomas	G.Hill	Blood	x				
14 01	Fall	Affected	G.Thomas	G.Hill	Blood	x				

Sample Name	Breed	Phenotype	Collected by:	Extracted by:	Sample type:	Microsatellite mapping	GWAS	Fine mapping (including candidate gene interrogation)	454 Re-sequencing (including interrogation of variants)	Global Transcriptionomics
09_00	Fall	Affected	G.Thomas	G.Hill	Blood	x	x*	x	x	
01_03	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
01_02	Fall	Affected	G.Thomas	G.Hill	Blood	x	x*	x	x	
30_02	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
86	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
27_01	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
28_99	Fall	Affected	G.Thomas	Author	Blood	x	x	x	x	
30_99	Fall	Affected	G.Thomas	Author	Blood	x	x	x	x	
38_99	Fall	Affected	G.Thomas	Author	Blood	x	x	x	x	
D2	Fall	Affected	G.Thomas	Author	Tissue	x	x*	x	x	
2F	Fall	Affected	G.Thomas	Author	Tissue	x	x*	x	x	
38_02	Fall	Affected	G.Thomas	Author	Blood	x	x	x	x	
31_02	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
48_98	Fall	Affected	G.Thomas	G.Hill	Tissue	x	x	x	x	
15_03	Fall	Affected	G.Thomas	G.Hill	Blood	x	x	x	x	
13_03	Fall	Affected	G.Thomas	G.Hill	Blood	x*	x	x	x	
03_09	Fall	Affected	Author	Author	Tissue Bone Marrow				x	x
04_09	Fall	Affected	Author	Author	Tissue Bone Marrow				x	x
06_09	Fall	Affected	Author	Author	Tissue				x	x
14_09	Fall	Affected	G.Thomas	Author	Blood	x	x	x	x	
07_01	Fall	Affected	Author	Author	Blood	x	x	x	x	
482	Fall	Affected	Author	Author	Fur	x	x	x	x	
517	Fall	Affected	Author	Author	Fur	x	x	x	x	
537	Fall	Affected	Author	Author	Fur	x	x	x	x	
538	Fall	Affected	Author	Author	Fur	x	x	x	x	
575	Fall	Affected	Author	Author	Fur	x	x	x	x	
576	Fall	Affected	Author	Author	Fur	x	x	x	x	
607	Fall	Affected	Author	Author	Fur	x	x	x	x	
612	Fall	Affected	Author	Author	Fur	x	x	x	x	
455	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
07_08	Winter	Carrier	Author	Author	Blood				x	
10_08	Winter	Carrier	Author	Author	Blood				x	
98_00	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
15_01	Fall	Carrier	G.Thomas	Author	Blood	x	x	x	x	
18_02	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
465	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
02_02	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
18_03	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
24_01	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
23_01	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
17_03	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
145_00	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
10_03	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
09_03	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
11_08	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
65_02	Fall	Carrier	G.Thomas	Author	Blood	x	x	x	x	
64_02	Fall	Carrier	G.Thomas	G.Hill	Blood	x*	x	x	x	
32_02	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
17_01	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
02_00	Fall	Carrier	G.Thomas	G.Hill	Blood	x	x	x	x	
14_08	Fall	Carrier	Author	Author	Blood	x	x	x	x	
15_08	Fall	Carrier	Author	Author	Blood	x	x	x	x	
13_08	Repeat sample of 465	Carrier	Author	Author	Blood	x	x	x	x	
07_09	Fall	Carrier	Author	Author	Blood	x	x	x	x	
08_09	Fall	Carrier	Author	Author	Blood	x	x	x	x	
09_09	Fall	Carrier	Author	Author	Blood	x	x	x	x	
26_09	Fall	Carrier	Author	Author	Blood	x	x	x	x	

APPENDIX 2**DNA EXTRACTION PROTOCOL - BLOOD (3.0mL-10mL)**

DNA was isolated from whole blood using the BLOOD AND CELL CULTURE BACC3 Nucleon extraction kit (SL 8512), as per manufacturer's instructions, except a maximum of 7.5mL of whole blood was used, with the following amendments to the protocol:

Stage 1: Cell preparation from whole blood

- 2: As per manufacturer's instructions except, the total volume was made up to 25mL with Nucleon A reagent.
3. Sample inverted 10 times with the Nucleon A reagent and centrifuged at 13,000g for 5 minutes.
4. Pour off the Nucleon A reagent, retaining 10% of the total volume from step 2 (maximum 2.5 mL).
5. Vortex vigorously to disturb the pellet and centrifuge at 13,000g for 5 minutes.
6. Pour off the Nucleon A reagent, retaining only the pellet.

Stage 3: Deproteinisation

1. As per manufacturer's instructions but the sodium perchlorate was increased to 750µl and inverted 25 times.
2. As per manufacturer's instructions except the sample was inverted for 3 minutes by hand.
3. Nucleon resin was omitted from the protocol and the sample centrifuged at 13,000g for 3 minutes.

Stage 4: DNA precipitation

1. As per manufacturer's instructions, except the mid-phase is not brown as Nucleon resin omitted.
2. Omitted stage 2.
3. As per manufacturer's instructions except 5ml of cold absolute ethanol was added to precipitate the DNA.

DNA Washing

1. The spool of DNA was 'hooked' from the ethanol using a glass hook and added to a 15ml falcon tube containing 2mL of 70% ethanol at 4-6 °C.
2. Step omitted. In place the DNA spool in the 70% ethanol was centrifuged at 4000g for 2 minutes.
3. As per manufacturer's instructions except the DNA was re-suspended in TE Buffer (200 – 750mL).

APPENDIX 4

MICROSATELLITE MARKERS USED FOR LINKAGE ANALYSIS AND
HOMOZYGOSITY MAPPING

Count	Marker Name	Chrom.	Position (Mb) on equine assembly 2 (EquCab2) as determined using ENSEMBL Blast tool	Reference	Primers for unpublished microsatellite markers (F = forward, R = reverse)
1	HLM005	1	1.63	Vega-Pla et al. 1996	
2	HP27	1	14.34	L. Skow, pers. comm.	F: TGTTCAATCAACCAATCTGCCC R: AAACCCCTCCACTACCCCATTC
3	ASB041	1	18.27	Irvin et al. 1998	
4	HMS059	1	26.31	Milenkovic et al., 2005	
5	AHT026	1	41.04	Swinburne et al., 2000b	
6	COR100	1	50.78	Tallmadge et al. 1999b	
7	UCDEQ487	1	66.49	Exzellston-Stott et al. 1997	
8	TKY015	1	71.24	Hirota et al. 2001	
9	SGCV002	1	76.35	Godard et al. 1997	
10	AHT040	1	89.89	Swinburne et al., 2000b	
11	UM041	1	91.61	Swinburne et al. 2000a	
12	LEX077	1	95.69	Bailey et al., 2000	
13	ASB008	1	99.98	Braun et al. 1997	
14	ICA043	1	110.28	Swinburne et al. 2000a	
15	ICA025	1	117.76	Swinburne et al. 2000a	
16	AHT058	1	127.64	Swinburne et al., 2003	
17	TKY295	1	133.15	Tozaki et al. 2000b	
18	ICA016	1	137.18	Swinburne et al. 2000a	
19	COR006	1	161.49	Hoopman et al. 1999	
20	HMS007	1	162.38	Guerin et al. 1994	
21	COR053	1	182.81	Ruth et al. 1999	
22	ASB018	2	5.26	Braun et al. 1997	
23	TKY384	2	9	Tozaki et al. 2001	
24	TKY003	2	22.74	Tozaki et al. 1995	
25	AHT035	2	23.35	Swinburne et al., 2000b	
26	HMS054	2	28.46	Milenkovic et al., 2005	
27	AHT067	2	30.94	Swinburne et al., 2003	
28	HMS051	2	32.98	Milenkovic et al., 2005	
29	TKY340	2	39.27	Tozaki et al. 2001	
30	UM129	2	64.17	Mickelson et al., 2003	
31	A14	2	74.47	Marti et al., 1998	
32	NVHEQ224	2	76.04	K. Reed pers. comm.	F: GTGACATGGCCTTCTATCC R: CTAACCTGGCATTCCCTTTC
33	UM076	2	87.06	Roberts et al. 2000	
34	TKY798	2	93.96	Tozaki et al., 2004	
35	AHT064	2	100.81	Swinburne et al., 2003	
36	VHL123A	2	109.82	van Haeringen et al. 1998a	
37	COR026	2	117.18	Murphy et al. 1999	
38	COR028	3	11.07	Murphy et al. 1999	
39	COR033	3	13.47	Murphy et al. 1999	
40	AHT022	3	21.13	Swinburne et al. 1997	
41	AHT090	3	31.62	Swinburne et al., 2003	
42	LEX057	3	36.31	Cozle and Bailey, 1997	
43	ASB023	3	79.28	Irvin et al. 1998	
44	LEX007	3	86.98	Cozle et al. 1996a	
45	AHT097	3	99.04	Swinburne et al., 2003	
46	UM192	4	3.13	Mickelson et al., 2003	
47	HMS006	4	7.23	Guerin et al. 1994	
48	TKY223	4	8.65	Tozaki et al., 2000c	
49	LEX050	4	49.36	Cozle and Bailey, 1997	
50	TKY552	4	65.15	Tozaki et al., 2004	
51	TKY375	4	72.56	Tozaki et al. 2001	
52	TKY363	4	96.44	Tozaki et al. 2001	
53	AHT042	4	107.02	Swinburne et al., 2000b	
54	NVHEQ102	5	21.78	K. Reed pers. comm.	F: CAACTGGGCCTCAATCTTGG R: AGGGITGGGTCATCATCC
55	TKY271B	5	25.66	Kakoi et al. 2000	
56	HMS052	5	28.65	Milenkovic et al., 2005	

Count	Marker Name	Chrom.	Position (Mb) on equine assembly 2 (EquCab2) as determined using ENSEMBL Blast tool	Reference	Primers for unpublished microsatellite markers (F = forward, R = reverse)
215	AHT082	27	27.27	Swinburna et al. 2003	
216	VHL150	27	29.65	van Haeringen et al. 1998b	
217	COR017	27	35.28	Hoopman et al. 1999	
218	UM003	28	10.56	Meyer et al. 1997	
219	TKY320	28	25.54	Tozaki et al. 2000b	
220	UM166	28	30.19	Mickelson et al. 2003	
221	TKY299	28	33.87	Tozaki et al. 2000b	
222	TKY364	28	40.09	Tozaki et al. 2001	
223	UCDEQ425	28	43.09	Eggleston-Stott et al. 1997	
224	COR082	29	4.28	Tallmadge et al. 1999b	
225	TKY628	29	18.05	Tozaki et al. 2004	
226	ASB043	29	30.34	Irvin et al. 1998	
227	LEX025	30	2.04	Coopie et al. 1996c	
228	HTG027	30	7.29	Lindgren G. 2000	
229	UMNE530	30	11.73	Mickelson et al. 2004	
230	UCDEQ455	30	27.41	Eggleston-Stott et al. 1999	
231	TKY368	31	4.21	Tozaki et al. 2001	
232	TKY274	31	11.46	Tozaki et al. 2000a	
233	TKY278	31	21.42	Tozaki et al. 2000a	

APPENDIX 5**SOFT TISSUE EXTRACTION PROTOCOL**

DNA was isolated from soft tissue using the BLOOD AND CELL CULTURE BACC3 Nucleon extraction kit (SL 8512) using the following protocol. A minimum of 50mg and a maximum of 200mg of soft tissue was used for the extraction.

1. Tissue was sliced and placed in a 15mL falcon tube with 700 μ l of Nucleon reagent B and 20 μ l of proteinase K solution.
2. Incubate the sample at 55 °C for approximately 20hrs (until all the tissue has dissolved). Vortex at regular intervals to aid tissue digestion.
3. Add 270 μ l of sodium perchlorate to the sample from step 2 and vortex vigorously.
4. Add the sample from step 3 to a clean 15mL falcon tube containing 700 μ l chloroform. Invert 10 times and centrifuge at 4400 rpm for 3 minutes.
5. Transfer the upper aqueous phase to a clean 2mL eppendorf containing 2.5mL of absolute ethanol at 2-4 °C. Invert the tube slowly to precipitate the DNA and centrifuge at 4000 rpm for 10 minutes.
6. Discard the ethanol and add 1mL of 70% ethanol to the eppendorf containing the DNA pellet. Dislodge the pellet from the base of the eppendorf by flicking the tube and re-centrifuge at 4000 rpm for 1 minute.
7. Pour off the waste ethanol and dry the pellet at 37 °C (until all of the ethanol has evaporated).
8. Re-suspend the DNA in TE Buffer (approximately 750 μ l).

APPENDIX 7



**HETEROZYGOUS HIGH QUALITY DIFFERENCES DETECTED IN
SAMPLE 03_08**

Count	Position	Variant detected by re-sequencing (03_08)		Ambiguous ?	Validation result (Sanger sequencing of 03_08)	
1	29,851,435	G	A	unambiguous	-	
2	29,928,903	C	G	unambiguous	-	
3	30,053,291	C	T	yes	C	T
4	30,079,153	C	T	yes	T	T
5	30,079,186	T	G	yes	G	G
6	30,079,206	C	T	yes	T	T
7	30,079,251	G	A	yes	A	A
8	30,081,827	T	G	yes	G	G
9	30,082,378	A	G	yes	G	G
10	30,082,514	G	A	yes	A	A
11	30,082,539	C	A	yes	A	A
12	30,082,549	T	G	yes	G	G
13	30,082,582	A	G	yes	G	G
14	30,082,592	A	G	yes	G	G
15	30,082,609	A	G	yes	G	G
16	30,082,632	G	A	yes	A	A
17	30,082,681	C	A	yes	A	A
18	30,082,717	C	A	yes	A	A
19	30,082,730	T	G	yes	G	G
20	30,082,758	A	C	yes	C	C
21	30,082,785	C	T	yes	T	T
22	30,372,387	C	A	yes	C	A
23	30,372,577	G	A	yes	G	A
24	30,747,620	G	A	yes	G	A
25	30,747,628	G	C	yes	G	C
26	30,747,649	G	A	yes	G	A
27	30,747,762	T	C	yes	T	C
28	30,750,837	C	T	yes	C	T
29	30,752,752	T	C	yes	Failed	
30	30,756,965	T	C	yes	Failed	
31	30,770,212	A	G	yes	Failed	
32	30,770,586	T	C	yes	Failed	
33	30,771,445	C	T	unambiguous	-	
34	30,778,047	G	A	unambiguous	-	
35	30,787,772	C	A	unambiguous	-	
36	30,793,941	G	A	unambiguous	-	
37	30,806,449	C	A	unambiguous	-	
38	30,810,972	T	C	unambiguous	-	
39	30,812,809	T	G	unambiguous	-	
40	30,813,759	C	A	unambiguous	-	
41	30,813,831	C	T	unambiguous	-	
42	30,815,472	T	C	unambiguous	-	
43	30,815,529	G	A	unambiguous	-	
44	30,817,434	C	T	unambiguous	-	

GENOTYPING RESULTS FOR THE HETEROZYGOUS SNPs DETECTED IN 03_08 IN 36 ADDITIONAL FIS-AFFECTED INDIVIDUALS FOR DEFINING THE BOUNDARIES OF THE FIS CRITICAL REGION

SNP Position	03_08	86	482	517	537	538	575	576	607	612	01_02	01_08	02_08	03_03	03_09	04_09	05_08	06_08	06_09	07_01	09_00	09_08	14_01	14_09	15_09	16_01	22_01	28_99	30_02	30_99	31_03	32_98	38_99	39_98	43_02	48_98	97_00							
30,053,291	C	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	T				
30,372,387	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	C	A			
30,372,577	G	A	G	G	G	A	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	A		
30,747,620	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
30,747,628	G	C	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
30,747,762	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
30,750,837	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

APPENDIX 8**CYTOLOGY REPORT FOR BONE MARROW ASPIRATES**

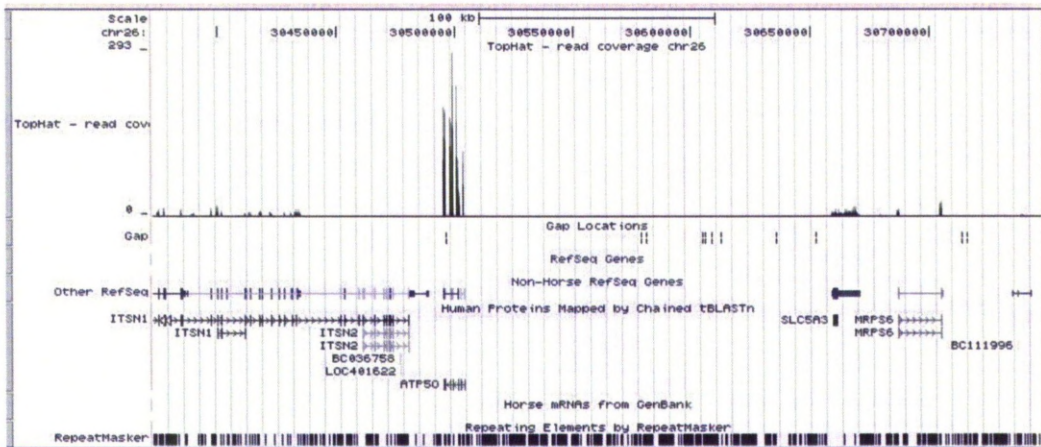
		Animal Health Trust Diagnostic Laboratory Services Diagnostic Services, Laxwades Park, Kentford, Newmarket, Suffolk CB8 7UU T: (01638) 552993 F: (01638) 555643 e: diagnostics@ahtr.org.uk			
To L Fox-Clipsham Genetics Student - CPM Section Head: J Swinburne AHT		Animal FELL PONY FOALS Owner Unknown Species Equine Breed Fell Gender Unknown Age Unknown Client Ref	Lab No PC491563 Type FNA (slides) Sampled 01/08/2009 Received 01/08/2009 Authorised 06/10/2009 Sample Ref ORDER No C09-20568		
Pathology - Cytology					
Bone Marrow Aspirate	PC 491563 Equine				
Bone Marrow Aspirates from healthy fell pony foals					
Description:					
15-09 Cellularity is low. The aspirate is heavily haemodiluted with abundant background erythrocytes, scattered platelets and platelet clumps. No flecks of bone marrow are seen and there is no cell monolayer. Scattered throughout the background blood are band and segmented neutrophils, eosinophils and monocytes. No megakaryocytes or blast cells are seen. A myeloid to erythroid ratio is approximately 1:1.7.					
16-09 The appearance of the aspirate is similar to that described above. Cellularity is very low with no flecks and numerous erythrocytes and platelet clumps. There are many segmented neutrophils and metarubricytes present with a few eosinophils. Megakaryocytes are not seen					
17-09 Cellularity is low, without flecks and there is marked haemodilution. Occasional metamyelocytes are seen with more frequent band and segmented neutrophils. Occasional eosinophils and precursors are seen. There are metarubricytes and an occasional rubriblast. A myeloid to erythroid ratio is approximately 1. Megakaryocytes are not seen					
18-09 Cellularity is low; flecks are not seen. There is abundant blood and platelets present. Segmented neutrophils are seen and there are a few eosinophils. Metarubricytes are present and the M:E ratio is approximately 1. Megakaryocytes are not seen.					
19-09 Cellularity is low with no flecks and cell preservation is poor. The nucleated cells appear pyknotic. Neutrophils, eosinophils and erythroid precursors can be identified.					
20-09 Cellularity is low, but higher than in previous samples. Again there are no flecks or cell monolayer and cell preservation is poor with pyknotic cells. Rubricytes and metarubricytes are seen. There are neutrophils and the M:E ratio is approximately 1:1.8. Megakaryocytes are not seen					
21-09 Cellularity is low, but again higher than previous. There are no flecks but a single megakaryocyte is seen. There are occasional metamyelocytes, band and segmented neutrophils and eosinophils. Occasional lymphocytes are seen and there are metarubricytes and rubricytes but there is no complete maturation sequence of either line; blasts are not seen. A M:E ratio is approximately 1.					
Comment Regrettably the bone marrow aspirates from 15-09, 16-09, 17-09, 18-09, and 19-09 are unlikely to be					
				Printed: 06/10/2009	Page 1 of 2

<p>To L Fox-Clipsham Genetics Student - CPM Section Head: J Swinburne AHT</p>	<p>Animal FELL PONY FOALS Owner Unknown Species Equine Breed Fell Gender Unknown Age Unknown Client Ref</p>	<p>Lab No PC491563 Type FNA (slides) Sampled 01/08/2009 Received 01/08/2009 Authorised 06/10/2009 Sample Ref ORDER No C09-20568</p>
<p>representative. Flecks of bone marrow have not been harvested and there is marked dilution from the peripheral blood, therefore it is not possible to perform a full 500 cell differential count. The M:E ratio is not likely to be fully meaningful: Many of the segmented neutrophils may be derived from the peripheral blood rather than the bone marrow. There are cells from both the myeloid and erythroid lineages present but regrettably re-sampling would be necessary to allow further evaluation.</p> <p>The aspirates from 20-09 and 21-09 are of higher cellularity and again cells from both myeloid and erythroid lines can be appreciated in both samples, however complete maturation sequences are not seen and so a 500 cell differential would not be appropriate.</p> <p>Please do not hesitate to call to discuss these control group ponies if we can be of help. It is regrettable that the samples preclude further evaluation.</p> <p>Case read at Cambridge University Vet School by: Sarah Putwain MA VetMB MRCVS, Resident in Clinical Pathology</p> <p style="text-align: center;">Alison Haslam</p>		

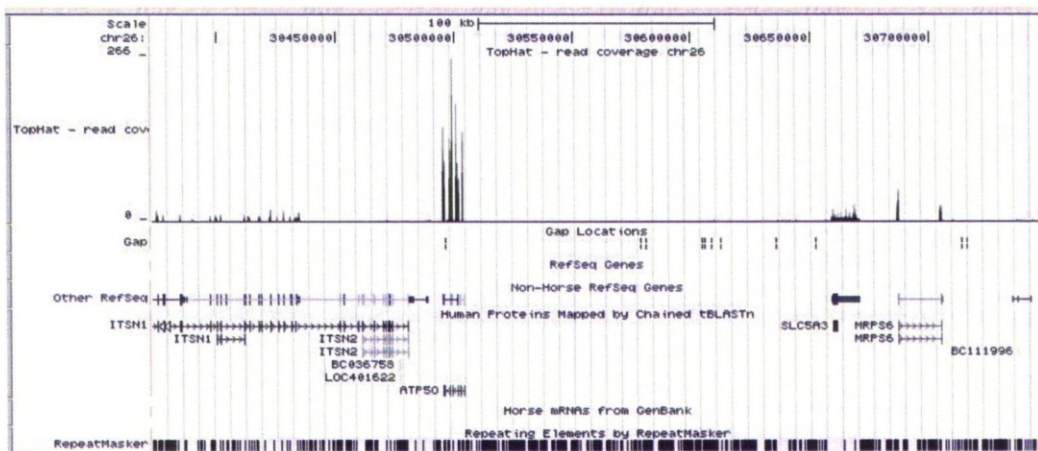
APPENDIX 9

UCSC ALIGNMENTS FROM RNA-SEQ EXPERIMENT

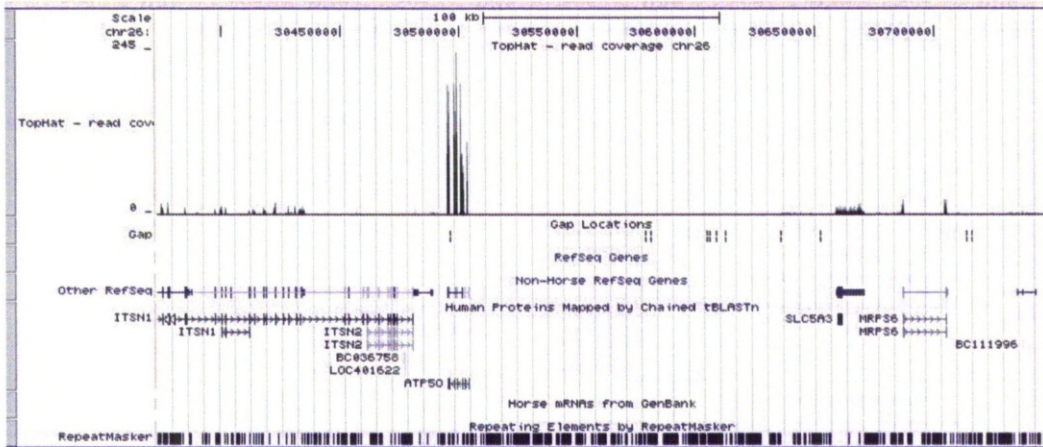
Sample 01_08 – Affected:



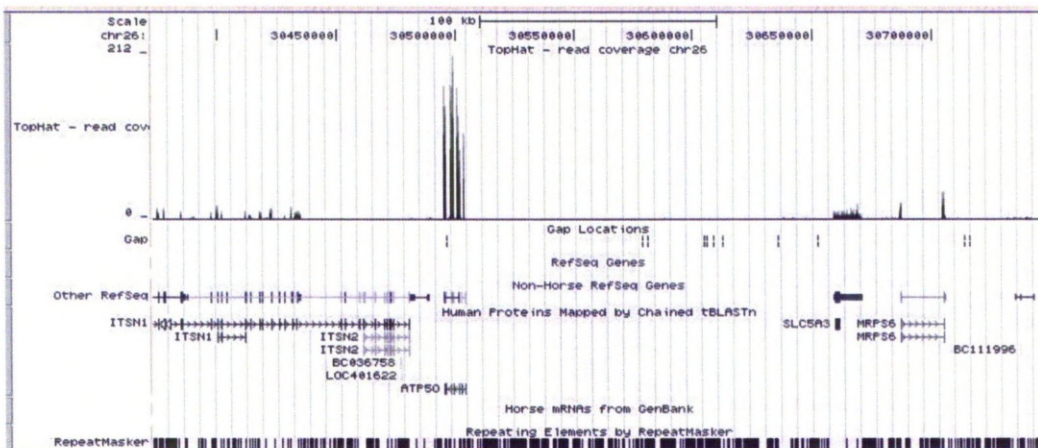
Sample 03_09 – Affected:



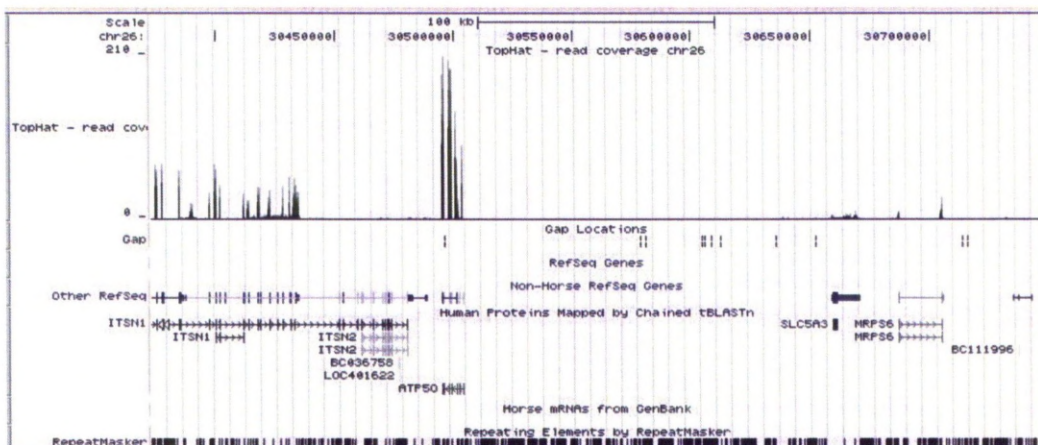
Sample 04_09 – Affected:



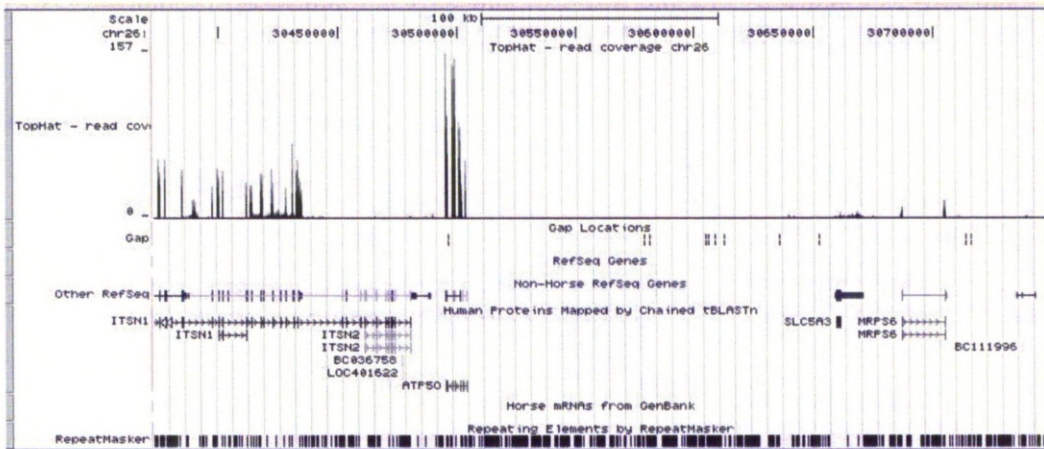
Sample 06_09 – Affected:



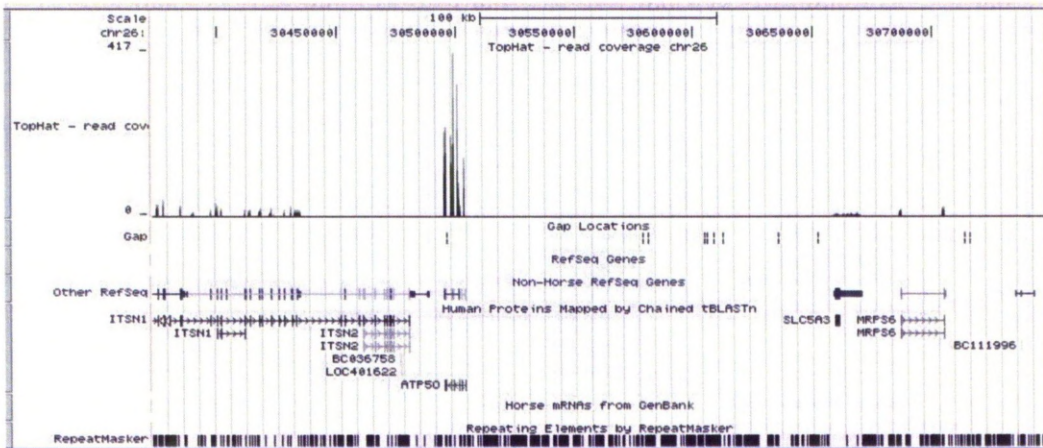
Sample 20_09 – Control:



Sample 21_09 – Control:



Sample 04_08 – Control sample which was euthanized due to clinical signs associated with colic. This sample as excluded from analysis:



APPENDIX 10

TOP 50 NETWORKS WHICH WERE IDENTIFIED AS SIGNIFICANTLY ALTERED IN FIS-AFFECTED ANIMALS

ID	Genes in Network	Score	Focus Molecule	Top Functions
1	AGTPBP1, BACH2 (includes EG-60468), CALML5, CDK3RAP2, CNTNAP1, CREB3L1, CRYBA4, CTH, DST, Dynein, EPB41, ERK, ERMAP, ENPH5, GCLC, GCLM, HMBS, LIMA1, MAF, MAFB, MAFG, NFE2, NUKA1, NYNRIN, PFAS, RAB27B, SPRED1, SYTL4, SYTL5, TBXAS1, TH1 Cytokines, TPT1 (includes EG-7178), TUBB, TUBB1, USF2	39	32	Drug Metabolism, Amino Acid Metabolism, Small Molecule Biochemistry
2	ABCB7, ADAMTS2, ARHGAP33, ATP6V0A1, BRPF3, CD226, DUB, DYNLT1, EPB41L2, EPB41L3, FAHD1, FCGRIA2A3A, FECH, FYN, HEXA, LAMA, MME, MMRN1, MPP2, MRV11, NARF, NRP1, peptidase, TMPO, TP53BP2, TRIM23, TROAP, USP1, USP4, USP7, USP25, USP35, USP38, USP42, ZMYND3	39	32	Post-Translational Modification, DNA Replication, Recombination, and Repair, Molecular Transport
3	ALAD, ASPM, BTG2, CSORF22, CDC80, CCNE2, CDCA4, CDK2-CyclinE, CENPF, CHMP2B, CMAS, Cyclin E, E2f, E2F2, ERBB2, ESPL1, FRRS1, HTRA1, IGI, MKD3, MYBL2, MYL9 (includes EG-10398), P4HA2, PRR11, PTRF, RAD21, RNF149, RRM2, SDPR, SMOCC2, SMTN, SRPX, SS18L2, TN33, WDR76	39	32	Cell Cycle, Embryonic Development, Cancer
4	AFAP1, Alpha catenin, ARVCF, CA2, Cadherin, CELSR2, CTNNA1, DDR2, EPB42, GNS, JAM, JAM2, KIF1C, KIRREL, LMO7, LPHN2, MLLT4, NCAPD3, NCAPG2, NCAPH2, Nectin, PAQR7, PARD3, PTFN21, PTPRG, SHANK3, SKAP1, SLC22A16, SMC4, SPTAN1, SRC, SSBP3, TALI, TCF7, TJP1	37	31	Cellular Assembly and Organization, DNA Replication, Recombination, and Repair, Cell Cycle
5	CD5L, COL8A1, DYRK1B, EFEMP1, EFEMP2, EGF7, EMILIN1, FAMI15A, FBLN5, FBN1, FBN2 (includes EG-2201), FSTL1, FGF1B, HSPG2 (includes EG-9339), LAMA2, LAMB1, LAMC3, Laminin1, Laminin2, Ltbp, LTBP1, LTBP2, LTBP3, MEGF8, NES, NID1, NID2, P38 MAPK, PLA2G7, Pp2c, PPM1D, PRKAA, RGS5, SGTB, TENC1	33	29	Genetic Disorder, Ophthalmic Disease, Inflammatory Disease
6	19S proteasome, 20S proteasome, ADRM1, ATPase, BCL2L1, BFAR, BNIP3L, CHRAC1, CLSPN, DNA-directed DNA polymerase, DNNT, GPR182, KIAA0101, LGI1, MYBBP1A, PCNA, POLD2, POLH, PPP2R4, PSMC4, PSMC8, PSMC14, RAD9A, RFC3, RPA, RPA2, SMARCA1, SYCAIP, TMBIM6, UBE2, UBE2C, UBE2J1, UBE2S, VDAC3, WRN (includes EG-7486)	33	29	DNA Replication, Recombination, and Repair, Cellular Growth and Proliferation, Nucleic Acid Metabolism
7	alcohol group acceptor phosphotransferase, AMOTL2, ANGPT4, ANGPTL2, ANGPTL4, CALCOO1, CDKN3, DUOX1, EGLN3, GNG11, GPM2, HIPK1, HMDLR, Importin alpha, Importin alpha/beta, Importin beta, KIF11, KIF15, KIF22, KPN1A1, LY75, MAP2K3, NDC80, NEK2, PLK1, PMF1, SGK1, SLBP, SPC24, SSTR1, STIL, TPK2, TTK, Vegf, ZWNT (includes EG-11130)	32	30	Cell Cycle, Cellular Assembly and Organization, DNA Replication, Recombination, and Repair
8	ADAMTS9, ANK1, BCR, BLK, CSAR1, CD19, CD22, CD79A, CD79B, ENPPI1, ENPPI2, ENPPI5, ETS, G0S2, HDGF, IgD, KLF3, MFHAS1, NFE2 (complex), NMRAL1, nucleotide diphosphatase, PAK5, PELI3, PIM2 (includes EG-11040), POU2AF1, PPARα-RXRα, RHAG, RJK3, SLC11A2, SPTA1, SPTB, TNFRSF13C, tyrosine kinase, UNC5CL, ZFAND6	31	28	Small Molecule Biochemistry, Cellular Development, Connective Tissue Disorders
9	Akt, APC-FZR1, AURKB, BMP5, BUB1, CCNF, CDC20, CDCA8, CENPA, CKAP2, COL4A2, DGAT2, E3 co-factor, ELOVL6, FBXO5, FKHR, FOXM1, Foxo, FOXO3, FOXO4, FZR1, HJURP, INCENP, LCN2, LXXN, MCOM, MTORC2, NUPR1, PPM1A, RNF10, SGOL1, SGOL2, SULF1, UBE2O, VLDL	30	28	Cell Cycle, Cellular Assembly and Organization, Embryonic Development
10	Alpha actin, Asp2/3, CALD1, CD58, CIT, CSTA, CTSH, CYGB, ECT2, EHD3, EPHX2, FAM46C (includes EG-54855), G-Actin, JUN, KALRN, KIF14, KIF23, KIF3C, MEF2, Mhc ii (family), Myosin, N4BP1, NFATC4, PAM, PARVA, PLK4, PRCL, PRSS23, PTX3, PDXN, RACGAP1, SDK1, TMOD1, TPM2, Tropomyosin	30	28	Cell Cycle, Cellular Movement, Cellular Compromise
11	ABCC3, ADA, AHSF, BLCAP, CLDN15, CYP1B1, DEPD1, ENDOD1, EPB41L4B, FAMI129B, FAM57A, GAS2L1, GCH1, Growth hormone, H3F3B, Histone h3, Histone h4, IRS1, KANK1, KIAA1598, LEPR, LIMCH1, LRRCSA, MIR1, MIR124, MXD4, NME4, ODZ3, P110, PAPSS2, PXR ligand-PXR-Retinoid acid-RXRα, SAA, SH3PXD2B, SNAI2, TMBIM1	29	27	Molecular Transport, Nervous System Development and Function, Organ Morphology
12	26S Proteasome, AMFR, BARD1, BIRC2, BTRC, Calcineurin protein(s), Calmodulin, CaMKII, CCNA2, CCND3, CCNE1, CDC6, CDC25B, Cdc25B C, CDK1, CDK2, CDR2, CHEK2, Cyclin A, Cyclin B, DLG1, DLGAP5, E3 RING, EPOR, EZH2, FAF1, MAF1, MDC1, MPP7, PER1, RBX1 (includes EG-9978), RNF5 (includes EG-6048), SERTAD2, SPA17, WDR26	29	27	Cell Cycle, Cell Death, DNA Replication, Recombination, and Repair
13	Cadherin (E.N.P.VE), CCR9, CDH5, DLL4, DNER, Eph dimer, FLT4, GREM1, JAG1, MERTK, Neuropilin, Notch, NOTCH3, NOTCH4, NRP1, NRP2, NRP2, PI3K, Plexin A, PLXDC2, PLXNA1, PLXNA2, PLXNA3, PLXND1, SDC2, Secretase gamma, Sema5, SEMA3A, SEMA3G, SEMA6D, SLIT2, STRA13, TLR2/3/4/9, TREM2, VASN	27	26	Cell-To-Cell Signaling and Interaction, Cellular Assembly and Organization, Cellular Movement
14	ACAC, ADIPOQ, APBB2, APP, Aspartyl Protease, BACE2, CLSTN3, COLEC12, CPT1, CTSE, EXT1, GAS6, Icam, ICAM2, ICAM5, ICAM4 (includes EG-3386), Integrin alpha 5 beta 1, LDB2, MAP1A, MAP1B, MAP1LC3B, MARCO, NLRG, PHEN, PITRMI, Plasminogen Activator, PLENI, PTPRD, Rab5, ROR2, SCARA3, SCARA5, SCAVENGER RECEPTOR CLASS A, SPON1, Vta-4	25	26	Cellular Compromise, Cellular Assembly and Organization, Cell Morphology
15	C1q, C1QB, C1QC, C1R, C1S, Collagen type IV, Elastase, GMFR, hydrolase, Igm, Immunoglobulin, LIPG, MASP1, MICAL2, Mmp, MDP19, NCEH1, NLN, NTSE, PAMR1, PCSK5, PCSK6, Pdgf, SDC4, SERPINA6, SERPING1, SLC25A37, Smaad, SORBS3, STAB1, STX12, TCF/LEF, TCF7L1, Trypsin, WASF3	24	24	Dermatological Diseases and Conditions, Genetic Disorder, Immunological Disease
16	Acid Phosphatase, ACP2, ACP5, AKAP12, AXL, Cbp, CSF1, CNCL12, DYRK3, ENCL, Erm, Estrogen Receptor, FKBP5, FKBP10, FSH, GAD1, KCTD15, KLF10, Lh, LOXL2, LZTR1, NFYC, p85 (pic3ε), peptidylprolyl isomerase, PPLI3, PRG2, RHOB, SLU7, Smaad1/5/8, SNAI1, Thymidine Kinase, TK1, TNFRSF11A, Ubiquitin, ZNF423	24	24	Skeletal and Muscular System Development and Function, Tissue Development, Connective Tissue Development and Function
17	ACHE, Alpha Actinin, AQP4, BOC, Calpain, CAPNS, CAPN11, CBX7, CDH1, CLIC5, CTSC, ENPEP, EPX, FCGBP (includes EG-8857), FCGRT, Fibrinogen, Gm-csf, GPT, GTSF1, Ige, IgG, IgG1, IGHE, IL12 (complex), KNG1 (includes EG-9827), LAMP3, MARK3, Mek, MMRN2, PDM1, PLS3, Ppp2c, PTPRU, Rak, STOM	23	24	Cell Death, Dermatological Diseases and Conditions, Immunological Disease
18	AFF1, Basal transcriptional machinery, BRD8, Cbp/p300, CCNB1, CDKN1B, CNTN2, Creatine Kinase, CREBBP, CTCF, EP300, GTF2B, HIST1H2AD, HIST2H4A, HISTONE, MED12L, Mediator, NCOA, NPAS2, PDK4, PKMYT1, PPARGC1A, RNA polymerase II, RNF123 (includes EG-63891), SLC20A1, Smaad2/3, TEAD1, TEAD2, Thyroid hormone receptor, TIP60, UBAC1, UBE2B, VitaminD3-VDR-RXR, WWTR1, YAP1	23	24	Cell Cycle, Gene Expression, Infection Mechanism
19	AEBP1, ANK3, ARHGAP6, ARHGAP29, CNTF, CNTF receptor, CNTFR, ERBB, ETV5, FBLN1, Fgf, FGF13, FGF23, Fgfr, FGFR4, Hspg, IL6ST, IAK, KIAA0467, LIFR, Mapk, MAZ, MPG, MPL, OSMR, PLC gamma, PTFN13, RAD23A, Rho gdi, RhoGap, Sbc, SLC1A3, SFRY1, STAT, UBL7	22	23	Cellular Development, Nervous System Development and Function, Tissue Morphology

ID	Genes in Network	Score	Focus Molecules	Top Functions
20	Adaptor protein 2, Ap1, Ap2 alpha, Arf, ASF1A, Beta Arrestin, Clathrin, CTSL2, DAB2, DGCR8, DNAAJ4, DNABJ4, DNABJ2 (includes EG-3300), DNACJ6, DNACJ9, Dynamin, EIF2AK1, HERC1, HIST4H4, Hsp90, Hsp22/Hsp40/Hsp90, ITSN1, MYO6, PADI4, Phosphatidylinositol4,5 kinase, Pias, PIP3K1B, PIP3KL1, PLK, RNASEN, SCAMP1, SNAP91, SNCG, STAB2, STON2	22	23	Cellular Function and Maintenance, Cellular Assembly and Organization, Cell Morphology
21	Casein, CD33, Dgk, DGKQ, EFN81, EGF, EGFR/PDGFR/IGFR, Ephb, Gap, GGT5, growth factor receptor, INPP5K, Integrin alpha 4 beta 1, NCK, PBK, Pc-Pic, PDGFR, Pdgfr, PDGFRA, PDGFRB, PI3K p85, PPAP2B, PSMF1, PYGM, RB1CC1, RGL1, SDC3, Sea, SP4, SPG20, SPHK1, ULK1, VAV, VRK1, YES1	21	22	Cancer, Skeletal and Muscular Disorders, Cellular Movement
22	ADAM12, ADAM15, ADAM19, ANGPT1, Cavosolin, EPB49, EPHA3, EPHA4, EPHA8, EPHB4, FERMT2, FZD4, GPC3, Integrin, Integrin alpha 3 beta 1, Integrin alpha 6 beta 1, Integrin, ITGA2, ITGA6, ITGA7, ITGA9, ITGA11, ITGA2B (includes EG-3674), ITGA8 (includes EG-8516), Ink, JUN/JUNB/JUND, Laminin, MAP4K5, Metalloprotease, PROCR, Rap1, RELN, SPAG5, Talin, TSPAN	21	22	Cell-To-Cell Signaling and Interaction, Cellular Assembly and Organization, Tissue Development
23	Aconitase, Actin, ANGPTL4, B4GALT6, CCR2, CYSLTR1, DLL4, DPEP2, GEM1, GJA4, GPR155, HLX, HPSE, IGHE, IKBKG, IL1B, IRG1, KIAA0101, KIF18A, LGTN (includes EG-1939), MND1, NCAFP3, NUSAP1, PADI4, PPP1CA, PQLC3, PWP2, PYHIN1 (includes EG-149628), RAD51, RAD51AP1, RAD54L, RECQ4, SHOX, TCN2, TOM1L1, TP53, TTC5, UBE2S	21	22	Cell-mediated Immune Response, Cellular Movement, Hematological System Development and Function
24	ANLN, APOBEC1, APOBEC2, ASPM, ATAD2, CDCA2, CDKN1A, CLCA2 (includes EG-9635), CPON, DLGAP5, DSE, E2F8, FAM3C, HURP, KIAA0101, KIF18A, LGTN (includes EG-1939), MND1, NCAFP3, NUSAP1, PADI4, PPP1CA, PQLC3, PWP2, PYHIN1 (includes EG-149628), RAD51, RAD51AP1, RAD54L, RECQ4, SHOX, TCN2, TOM1L1, TP53, TTC5, UBE2S	20	23	Cell Cycle, DNA Replication, Recombination, and Repair, Cellular Assembly and Organization
25	AKAP1, Ampa Receptor, ANTXR1, APC, C13ORF15, CCR7, COL6A1, COL6A2, COL6A3, Collagen Alpha1, DMP1, DUSP1, FAT1, GNRH, GPR56, HOMER1, HPSE, IgG2a, ITPR, JNK1/2, KAT2B, KDM6B, NCOA4, p160, p70 S6k, Pta, Pta catalytic subunit, PRKAR2B, PROM1, RAB13, Rar, T3-TR-RXR, Tgf beta, TGM2, TRPC6	19	21	Genetic Disorder, Skeletal and Muscular Disorders, Connective Tissue Disorders
26	ACSL3, ALOX5, ALOX12, ALOX15, APOE, CCL3L3, CH3L1, COL11A1, COL16A1, DPYSL3, DYRK2, HEBP1, HOKA2, HOKD3, HSD17B6, ITGA3, ITGA6, ITGB5, MEIS1, NR1H3, PARP, PBX1, PDZK1IP1, PF4, SGK1, SLC39A14, SLIT3, SVEP1, TGFB1, TPST2, TRIB2, TSPAN7, uric acid, WNK1, ZFYVE9	19	21	Lipid Metabolism, Small Molecule Biochemistry, Molecular Transport
27	AMMECR1, BCL6, CALU, CCR7, CD19, CD72, CD79B, CEP70, CIT, CLEC1B, EPB41L2, GNB2L1, GPC4, MAP4K3, MIR182 (includes EG-406958), MIRLET7B (includes EG-406884), MNTA3, NDN, NOL6, PCDH9, PTN, RCOR3 (includes EG-55758), RHOJ, SDC1, SDC3, SEC26, SLC27A6, SLC9A9 (includes EG-285195), SPI1, SPIB, TAF1D, TMEM86A, TUBA3C, ZNF426, ZNF532	19	21	Connective Tissue Disorders, Dermatological Diseases and Conditions, Genetic Disorder
28	ADD2, CD36, collagen, Collagen type I, Collagen type III, Collagen(s), Complement component 1, CYR61, DDR1, Fascin, FBLN1, Filamin, FSCN1, GP5, GP6, GPIIb-IIIa, Integrin alpha V beta 3, ITGB5, Kallikrein, Laminin b, LBR, MAST2, Mmr, MRC2, MYO10, PCOLCE, Pkc(s), Ptg, PLA2R1, S100A1, SERPINH1, SLC6A9, SP2, VWF, ZNF384	19	21	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Inflammatory Response
29	ADAMTS12, AMIGO3, CNPK2, CNTFR, CPEB4, DAD1, DDX60, DNABJ2 (includes EG-3300), DOLPP1, FAM100B, FBXO8, FANL3, FANL2 (includes EG-114793), FRYL, FURIN, GOLPH3, GRTP1, IFN1@, IFNA2, ISG20, Irf, ITH15, LTBP3, MIR125B1, MIR199A2, MIR206 (includes EG-406989), MDC, PCSK6, PDIK1L, PRPF40A, RCN3, SAMSN1, STAT4, TSPAN33, XBP1	19	20	Embryonic Development, Tissue Development, Post-Translational Modification
30	Alpha tubulin, ANLN, ASF1B, Beta Tubulin, CCNB2, Cdc2, Cdc25, CDK6, CDKN2C, CDKN2D, DYX1F, E2F1, E2F4, Gamma tubulin, Hdac, HDAC6, HDAC11, Ink4, MCM2, MCM5, MCM7, Mpf, NDN, NGF, p13-kinase, RAB11FIP3, Rb, SUV39H1, SYN1, TEK, TFPD2, TUBB3, TUBB4, TUBG2, Tubulin	17	23	Cancer, Neurological Disease, Dermatological Diseases and Conditions
31	ACO1, ALAS2, CLK3, CNKLR1, EIF2AK2, IDO1, IFIT1, IFITM5 (includes EG-387733), Ifn, IFN alpha/beta, IFN Beta, Ifn gamma, IFN TYPE 1, IKBKAP, Ikk (family), IL1, IL23, IL12 (family), IL15RA, Interferon Regulatory Factor, IRF, IRF2, IRF4, IRF7, IRG, KLF15, LTB, MDC2, NFkB (family), Oas, OAS3 (includes EG-4940), TAB3, TGFBR3, Tnf, V5IG4	17	20	Antigen Presentation, Hematological System Development and Function, Tissue Morphology
32	CDO1, CIR1, Dishevelled, ECE1, ESM1, F8, F13A1, FHL1, FRIZZLED, FZD1, FZD8, Glutathione peroxidase, GPR183, GPX1, GPX8, hCG, HSPB6, IGFBP4, Ldh, LDL, LEF1, LRP, LRP4, LRP6, Mlc, NDN, PDGF BB, PDLIM1, PKA, RND3, Reck, Sod, STAT5a/b, WIF1, Wnt	17	22	Embryonic Development, Organismal Development, Skeletal and Muscular System Development and Function
33	A2M, ABCA1, ABCB10, ACSL, ADAMTS7, APOE, ARHGEP12, CYP7A1, FASN, FXR ligand-FXR-Retinoic acid-RXRa, HDL, HDL-cholesterol, K Channel, LCAT, LXR ligand-LXR-Retinoic acid-RXRa, MCF2L, N-coor, NCOR-LXR-Oxysterol-RXR-9 cis RA, Neurotrophin, Nr1h, NR1H2, NR1H3, NTRK2, NTRK3, Plexin B, PLTP, PLXNB2, PLXNB3, Ptk, Ras, Rar, Sfc, SORCS1, SPOCK2, Trk Receptor	16	19	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
34	AKT3, ATP11A, ATP5C1, ATP5J2, CD320, CDK4/6, COCH, COL4A1, COL4A2, CRIM1, CYP7B1, ESCO2, FBN1, GAA, GDDP2, IQGAP3, MIR101, MIR122 (includes EG-406906), MIR210 (includes EG-406992), MYC, MYCN, PDGFC, RBL1, RBP1, RPL38 (includes EG-6169), RPS7, RPS20, RPS23, SDC2, SP9, TGFBR3, TPD52, TPRA1, TSPAN7, XRCC3 (includes EG-7517)	16	19	Embryonic Development, Tissue Morphology, Gene Expression
35	ABTB1, AKT3, APCDD1, beta-estradiol, BRPF1, BTG1, C17ORF63, CAD, CCNB2, CXSA, CYTH3, EXT2, FAM118A, FAM78A, FBN1, FDF1, GPR64, GPX3, IFNGR2, IFRD2, IGFBP6, KIAA0922, MIR23B (includes EG-407011), MIRLET7G (includes EG-406890), NRP1, PINK1, PTEN, SBF1, SP4, SUSD2, TGFBR3, TNRC6B, UGGT1, YPEL3, ZNF1	16	19	Cellular Development, Cancer, Cellular Growth and Proliferation
36	ALDH1A3, ALDH3A1, DRG2, FAM118B, FHL1, GLI1, GPX1, HPRT1, IFI16, IFIT1, IFITM1, IGFBP6, KCTD11, MAO, MCC, NAA10, NAT2, NKX2-5, OAS2, PAFSS2, PCDH18, PDGFRA, PITX2, PLA2G15, PNP, PUS7, RARRES1, retinoic acid, SLC5A3, SMOX, TBX6, TBX15, TUFM, UCP3, VGLL3	15	18	Genetic Disorder, Respiratory Disease, Cell-To-Cell Signaling and Interaction
37	AQP1, ASRGL1, C15ORF59, C5ORF30, CCDC116, CRCP, CYFIP1, DCAF8, DNPEP, EBF1, EHPB1L1, EIF4E, FASTK, GEMIN5, ISLR, KIF26B, MDF1 (includes EG-4188), NQO1, PAX5, POLR2E, POLR2H, POLR3A, POLR3D, POLR3E, POLR3H, PRDM5, PRDM5, PRDX5, RFX2, RHOG, RPL30, SCPEP1, SLC2A4, TINAGL1, TP53BP2, ZNF521	15	18	Cellular Development, Cell Cycle, Hematological System Development and Function
38	ABLIM1, ALS2CL, ANKS1A, ARHGAP11A, ARHGEP5, ARHGEP17, CDC42EP2, CGN, CHMP4C, COBL, DBNL, F Actin, FGD1, HAGH, HECW2, INPP4A, KCTD5, KIF23, LIG1, MARK1, N4BP1, NKD2, RAB5A, REEP4, SEMA6A, SFN, SHROOM3, SORBS1, SYNE2, TP73, TTYH2, UBC, USP5, USP6, USP8	15	18	Carbohydrate Metabolism, Lipid Metabolism, Small Molecule Biochemistry
39	ABCA8, ACCN4, ACTB, ARHGEP10L, ATF7IP, ATG9A, BNF, C19ORF55, CCT5, CENPL, CENPN, CNTN4, CTTNBP2, CTTNBP2NL, DYNLL1, FAM40A, FAM40B, HIST2H3A, HSPB3, KIAA1632, MASTL, MIRN340, MIRN349, MOBKL3, MT381, NIPSNAP3A, OLFML3, ONEUT1, PHOSPHO1, RRP12, SETDB1, STK25, TRAF3IP3, VVASA, YPEL5	15	18	Genetic Disorder, Hematological Disease, Digestive System Development and Function
40	ABCC9, ADCY, ADCY2, ADCY6, ATAD2, ATP, ATP1F1, CCL14, CYSLTR1, EYAI, EYAZ2, F2R, G protein, G protein alpha1, G-protein beta, GUCY, GUCYL1A5, Gsl2/13, IP6K1, KCNJ8, MYO1A, MYO1B, ORC5L, P2RY1, Pmca, Potassium channel, Pp2b, PTGER3, PTGIR, RAB31L1, RGS16, S1PR3, S1X1, SOX6, Voltage Gated Calcium Channel	15	25	Respiratory System Development and Function, Hematological System Development and Function, Inflammatory Response

ID	Genes in Network	Score	Focus Molecule #	Top Functions
41	ABCA6, ACADM, APOA1, APOA4, BTG2, CIQA, CDKN2A, CDYL, CHRDL1, DEK, EIF1B, FAM111A, FKBP3, FOXM1, GPT, HDAC1, HMG2, KANK2, LIPC, MKR45, NR2F2, palmitoleic acid, POLK, PPARA, PPARα-RXRα, PSRC1, RAD51AP1, RBP2, RECQL4, RRM2, TCF19, TP53INP1, UCHL5, YY1, ZNF236	15	18	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
42	AHSG, ARHGAP11A, ATG2B, BSDC1, C18, C9ORF100, CLDN2, CREB1, DCLRE1B, DEDD2, EZH1, GAS1, GBE1, GSTCD, HAO1, HHLA2, HNF1A, HNF4A, INSR, ITIH4, LRRRC17, LRRRC31, MCCC1, MIR214 (includes EG-406996), MRO, NAT10, NRD1, NNT2, PTPRG, SURF6, TMCO3, TMEM2, WNT10A, WNT3A, YPEL3	14	17	Cellular Development, Cellular Growth and Proliferation, Endocrine System Development and Function
43	ARMCS, BNIP3L, CADM3, CAMSAP1, EMP1, EWSR1, FBXO34, FEM1C, GFBP1L1, HOXD4, LEMD3, LHX9, MIER3, MIR17 (includes EG-406952), MIR181B2, MIR19A (includes EG-406979), MIRN101B, NCSTN, NIPBL, NT5C3, ODF2, PAPPA, PCDH19, PIR, PRUNE2, RHBDL1, RANND5A, SERBP1, SMAD9, SOX4, TNPO1, WDR26, ZEB2, ZFHX4, ZIC1	14	17	Genetic Disorder, Hematological Disease, Cellular Development
44	ADAMDE1, AICDA, ARG1, CASP1, CD101, CENPT, ELM03, ENPP2, GAPDH (includes EG-14433), GAPDH (includes EG-2597), GAPDH5, Ggt, GJA1, GJA5, glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), GPNMB, GPR44, HBNIP, HGS, IFNGR2, Iga, IL2, IL13, LRRRC16A, LTA4H, mannose, PDGFC, RAB39, RFFL, SLC43A3, TNFRSF13B, TPT1 (includes EG-7178), TRAK2, ZBTB32	14	17	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Hypersensitivity Response
45	ADAMTSL1, BCAM, BLVRB, C11ORF31, C2ORF24, CBR3, CHST8, COX6B1, DECR1, DHRS1, ERO1L, FADS1, FAM65A, GTPBP2, HAS1, LAMAS, LAMB2, LSR, ME2, MIRN328, MIRN330, MOGAT2, NDUFA7, oxidoreductase, PDIA5, phosphatidylserine, POR, PRNP, PTDSS2, RECQL4, SECISBP2, SEPW1, SLC38A4, VCL, YOD1	14	17	Amino Acid Metabolism, Molecular Transport, Small Molecule Biochemistry
46	ACP1, ACVRL1, Ahr-aryl hydrocarbon-Arnt, CD6, CD163, COL5A1, COL5A2, COL5A3, Cpla2, Cyclin D, ENaC, ERK1/2, Ear1-Ear1-estrogen-estrogen, F11R, GC-GCR dimer, Glucocorticoid-GCR, ID1, MTUS1, NFIA, NFIB, Nuclear factor 1, PBM1, Pdgf Ab, PDGF-AA, PDGF-CC, PDGFC, Pdgfrb, Pdgfrb, PRKAC, Sma2/3-Smad4, Stat3-Stat3, SWI-SNF, TFF2, TIF2-NCOA1-p300-PCAF-CBP, TSC2D3, TSPYL2	13	17	Gene Expression, Connective Tissue Disorders, Dermatological Diseases and Conditions
47	ACACA, AMPK, AURKA, C1QTNF6, CDK5R1, CK1, Cofilin, DCK, DDIT4, ERRF1, FADS1, FADS2, Glycogen synthase, Gal3, Histone H1, INADL, Insulin, MLXIPL, Na+K+ -ATPase, NCAPD2, NCAPG (includes EG-64151), NCAPH, NMDA Receptor, PDE5A, PEPCK, PLA2G16, PP1 protein complex group, PP1-C, PP2A, PPP1R15B, PPP1R1B, Proinsulin, PTPase, Pyruvate kinase, TBX21	13	19	Lipid Metabolism, Small Molecule Biochemistry, Cellular Assembly and Organization
48	14-3-3, ATG4A, BCL2L13, Caspase 3/7, CCNDBP1, CD3, CD5, CD8, CD52, CD3-TCR, CD3D, CD9E, CD3G, CD8A, DKK3, Fcer1, GABARAPL2 (includes EG-11345), GBA3, GRAP2, IKK (complex), Interferon alpha, LCK, MAP2K1/2, MHC Class II (complex), MS4A1, NFAT (complex), Nfat (family), Pak, PTPRCAP, SEC14L3, SRC, SYNPO2, TACC2, TCR, ZAP70	13	20	Immunological Disease, Cancer, Hematological Disease
49	ABCB6, Alp, APOA1, ARG1, C14ORF126, Caspase, Creb, CYC1, Cytochrome c, Cytochrome c oxidase, EGR1, Endothelin, Focal adhesion kinase, G protein beta gamma, GABPB2, Gpcr, Hemoglobin, HMG2, Hsp27, Hsp70, HSPA6, MAPK10, Nos, PLA2, PLCEL, Pld, PLD3, Pro-inflammatory Cytokine, Rac, Ras homolog, RNASEH2A, Sapk, SNCA, STRADB, THBS1	12	16	Cellular Assembly and Organization, Cell Death, Amino Acid Metabolism
50	ACPI, AKAP6, ASAP2, atypical protein kinase C, BPGM, BRD4, C16ORF5, CD2AP, CYP17B, EFS, EPHA2, EPHA4, Gef, GLTP, GRB2, growth factor receptor, LUC7L2, MICAL1, MICAL3 (includes EG-57553), MICALCL, MIR373, MIR185 (includes EG-406961), MIR295 (includes EG-100049713), MST1R, PHLPP2, Ptk, PTK2B, RAB1B (includes EG-81876), RASGEF1B, SKAP1, SKAP2, SLC20A1, STAP1, TMEN9B, YPEL4	12	16	Cell Morphology, Cellular Assembly and Organization, Cellular Function and Maintenance

APPENDIX 11

DETAILS OF PRIMERS DESIGNED FOR THIS INVESTIGATION

Primer Name	Forward	Reverse	Position
Candidate gene screening			
IFNAR1 Exon 1	GCGTATGGGTGCTAGGCATTG	ACCACGGACGACCGGGCGGTCT	29,977,612
IFNAR1 Exon 2	CAGTCTCCTTGCTGAGAGC	CCAAAACCTCTTCTGCTTCAA	29,982,564
IFNAR1 Exon 3	CCAGGACATTAAGTCAAGTGA	AGGGAAAGCAGAAAAATGAGG	29,986,896
IFNAR1 Exon 4	TGTAGGCTCCATGAGGAACCT	GCTGCACACGTAGTCTTGCTC	29,989,174
IFNAR1 Exon 5	TGAACATCTCTCCTCTGGAA Internal primer: CTCTCAAGCGGAGAAGT	CCTCTCTCTTTTGTGGATTC	29,989,359
IFNAR1 Exon 6	TGAGTGGCTGCGTGGTATTA	GATTAGCTCAATGGCATCTGG	29,990,445
IFNAR1 Exon 7	TGAGGGTGTGGACTGTTTTTC	GGGGGAAGGATGATAGCTAGA	29,993,791
IFNAR1 Exon 8	GTCCGTGTACAAGCGTCTAA Internal primer: GATACTGAAATACAAAGT	AACAGGCTGAGGGATCTGAAT	29,994,039
IFNAR1 Exon 9	CAACCAAGTGTGGGAGAGA	GCTTGGAGGTCTACTTTCGACT	29,996,430
IFNAR1 Exon 10	GCTGGCAGAAGTATTCCTGAG	CAACTATGACTGGGGCTTTGG	29,997,303
IFNAR1 Exon 11	TGATATGCTCCGAGCTGACTT	ACTTAGGAAACTGAAGCT	29,998,945
IFNAR2 Exon 1	CCAGAAGCTGGAGTAGGCTCT	AATCCAAACCATGAACACAGG	29,986,832
IFNAR2 Exon 2	TTTAGCACAAAGAAAGCCTCCA	CAAACAAAATGGGCAACAGA	29,898,013
IFNAR2 Exon 3	GACACTTGGGAGGGTTCTCT	TTGGCACCTTAAAGCTCTG	29,899,262
IFNAR2 Exon 4	AATCCAGGTGCTTCAGACAGA	CCGAAGCTCAGAGAAACAAAG	29,900,390
IFNAR2 Exon 5	TGCCTTAAGTAGGACCATAGGC	CAAAGTCCCACAAGGAATGAA	29,902,724
IFNAR2 Exon 6	GATGAATTTCTTGCCCAAGT	GCTTAGCGACAATCTTCTCA	29,906,294
IFNAR2 Exon 7	CAATGGCTACCGTATTGGTCA	TAGGGATGGAAAGAACCAGAA	29,909,928
IFNAR2 Exon 8	TATTGATGGTCGTCATCGTCA Internal primer: CAGGCCTGTGTCGGC	AATTTGCATAGATGGCGTCAC Internal primer: TAGGTCCACTGTCTCT	29,914,827
IFNGR2 Exon 1	AAAGGCGAGTTGTGTGAATTG	GAATGAGTGGGAGGTTGTGAA	30,064,527
IFNGR2 Exon 2	GCACAACCCTCTGAATACACTG	TGTGCTGGTACTTAGGGTGAG	30,070,613
IFNGR2 Exon 3	AACCTCCAGCTCTGTTTCTC	GCCTTGAAGAGGAGGATGAAC	30,074,442
IFNGR2 Exon 4	CTGTCCAGCGTAACTATTGG	TTCCAGAGATCTCAGCAGAG	30,078,441
IFNGR2 Exon 5	GCAGCATCTGTCCATCTTACC	TTTCTGCTTTTTCTCCCCAGT	30,079,031
IFNGR2 Exon 6	CGTTGGTGAACAGAAGAGGAA	AATAAGCTTGTCTCCCCGATG	30,082,228
IL10RB Exon 1	CAAGGTCCCTTTCTCTTTGG Internal primer: ACCGACTCTCGAGCAGG	CGATCAGTGAGCCTATGTGGT	29,918,620
IL10RB Exon 2	CAGCCTCCTCAGCACAGTTAC	GCAACACCCTCCTTTGTCTTT	29,919,512
IL10RB Exon 3	TTGAGGTTTACGGAGCACAG	TTCCCTAAGCTTCTTGTGCAG	29,923,894
IL10RB Exon 4	TAAAACCAGACTACCCTCCTT	TTGCAAGACCCTGACTGGAT	29,926,252
IL10RB Exon 5	AGCACCTGCTTCCTTCATCTT	GCGTCCACATACCACAATA	29,930,192
IL10RB Exon 6	CTACAGATCTCCCGCCTTC	GTGACCACAGACTTCCTTGAA	29,937,052
IL10RB Exon 7	CTGTCAGAGTGTATGGTCA	AAGTTCAGGGGCTGAAAA	29,940,780
Primers used for microsatellite fine-mapping (4 novel microsatellite markers. Only FPS1-3 were polymorphic and so used for fine mapping).			
FPS1	GGCAGTGGTTTTCTGTCTT	GTGAGAGGGCCTGTATGAAAA	29,848,542
FPS2	GAGTCGGTTGCTGCTGAATC	CAAAGACCCATTACCAGCTCA	29,947,425
FPS3	TGAAGCTTTCGGAAGTTAGG	ATTTGGAAGTGTCTCAGCAA	29,963,137
FPS4	TGGAGGTTCTATTGGTATGTGC	TTTGAATAAGATGGCTGGAC	30,009,051
Primers used for SNP fine-mapping (Of the 118 identified from the Broad Institute 62 were polymorphic in the Fell Pony and 15 were novel SNPs. Therefore 77 SNPs were used for fine-mapping.)			
BIEC567067B*	CAAATCCAGAGGTGGTTTCA	TAACACGCATGCTTCTACTG	29,808,990
BIEC567067A*	CAAATCCAGAGGTGGTTTCA	TAACACGCATGCTTCTACTG	29,809,082
BIEC567064	TGGTTGGTTAAGATGCACACA	ACCAGAACAGCCACAAAAACA	29,825,135
BIEC567063A*	ATTATCATCGAACCCCTCCAT	CTTGTGTATGCTAGCCACAT	29,825,158
BIEC567062	TGGCTCTGATGACAAGATGAA	TTAACCCTGCCAGTGTCCAAC	29,843,328
BIEC567061	CTGTGGGTCCTATGCCTGTT	TCTGAGGTGGTGGCTTAGAGA	29,846,751
BIEC567057	CCAGGCTGTATCCAGCTAAT	GGGTCATCACAATGTCATCCA	29,934,936
BIEC567056	CCAGGCTGTATCCAGCTAAT	TCACAATGTCATCCACAGTGC	29,934,963
BIEC567055	TTTCTTGGTGAATCAGGGTCA	GCCCAGCAATCAGTGTTAAT	29,934,998
BIEC567054	GGCTCAAGCAAACTACCTT	CCCCATCTGACCTCTGTTAT	29,935,124

BIEC567053	CTTCTCCCTTTGTGGAGACG	ACATGACCCTGATTACCAAG	29,935,162
BIEC567052	CCTGGGCATGTTTAAATGTCTC	TGTCCACACAAACAACCAAGA	29,935,257
BIEC567051	TCAGGGTCACATTGATCTGCT	CACAAAACAACCAAGAGCCATT	29,935,286
BIEC567050	GTGGTGATAGCTGCACAACAA	GCATCGTGATCGTCTAAAAA	29,940,175
BIEC567047	CTCATGTCCAAGCAGAAGCTC	GGCATTGGTAATGTCAGGAAA	29,944,756
BIEC576046	TTCATCTCATCCGAAAACATC	TTAGGCACAAACCCACTGAAC	29,960,925
BIEC567045B*	TTCATCTCATCCGAAAACATC	TCCAGGGAGAGATGGGTTAGT	29,960,940
BIEC567045A*	TTCATCTCATCCGAAAACATC	TCCAGGGAGAGATGGGTTAGT	29,960,960
BIEC567045	TTCATCTCATCCGAAAACATC	TCCAGGGAGAGATGGGTTAGT	29,960,987
BIEC567040	TTTTCTCAAAAAGCCCTAAAA	TAAGCCTCACATCCACAATCC	29,968,592
BIEC576036	TTTAGTTCAAGAAGGCAGAGCA	GATGTCCAGCATGGGGTATC	30,002,761
BIEC567035	TTTAGTTCAAGAAGGCAGAGCA	GATGTCCAGCATGGGGTATC	30,002,762
BIEC567022	CAGTTGAAGGCAGAAGAGTGG	GTCGGATGCAGATTTCACTGT	30,061,151
BIEC567017	CTGGAAACACAACGACAGACA	CGTTGGTGAACAGAAGAGGAA	30,082,319
BIEC567016	CTGGAAACACAACGACAGACA	CGTTGGTGAACAGAAGAGGAA	30,082,329
BIEC566975	TGAAGCGTTAACCAAGTCCAA	GCACAAAGGAAACTATCCCTCA	30,176,665
BIEC566971	GTCAATTTACTTCCCGTTTTG	AGGTCAGATGCCTCTTTAGTTCT	30,221,284
BIEC566970	CAAGCCTTATTTGTTTCCACAG	CTGTTTTGCTGCTAAAGTAGAGAAG	30,221,638
BIEC566969	TTGGAAACAGAAGGAATGTG	AGGAAATTAACACCCCTTCC	30,221,752
BIEC566966	CATTGCAGAGCGACAGAAGTT	TGCAATCCTACTCGATTGCTT	30,242,235
BIEC566956B*	ACTGAATCTGGGCTGCTGAAG	TTCCAGTTATTTCTGTGGTCTG	30,311,805
BIEC566956A*	ACTGAATCTGGGCTGCTGAAG	TTCCAGTTATTTCTGTGGTCTG	30,311,821
BIEC2-692793	GGGTTTGGAGAAAAGGAA	TGGAGCTTGACCACTTGGAA	30,368,089
BIEC566946	TCGGGATGTGTACAGATGTA	AAAAGTATGTGTCTCTGGGTTATGC	30,368,149
BIEC566945A*	TCGGGATGTGTACAGATGTA	AAAAGTATGTGTCTCTGGGTTATGC	30,368,158
BIEC566945	TCGGGATGTGTACAGATGTA	AAAAGTATGTGTCTCTGGGTTATGC	30,368,204
BIEC566936	CCACCAAGGTCTGTTTGTCTT	TTTAAGTAGCCCGTTAAGGAGA	30,396,161
BIEC566930	GGGGCATTAGCTATTCTCA	CCATGGAAGCATTCAATTAGC	30,432,335
BIEC566924	CAGACTATGACCCCGAGACAG	CTGGGGACATTTTCTTTAGC	30,507,003
BIEC566923	AAGAATGCTGGCAATCTGGTT	CACCCTTTGAGAACCTGTGAG	30,511,626
BIEC566922	GAGGGTGTGGCCACATCTAA	TGATGAAAAATGGCGGAATAG	30,511,708
BIEC566913	ACTCCGACGTAACACAACCAG	TTCAAGCAGCAAGGAACCTAT	30,642,736
BIEC2-692830	CAAACCTCAGATGCCAACCTA	GCAGGAGTGGGGTTTGATAAT	30,720,109
BIEC566898A*	AAACTTGAGCCAGCATCAGAA	CAGGTTGTTGGCTGGACTCT	30,736,819
BIEC566863A*	AGAGCTTTGCCTGGAAGAGAG	TCTATGACAGAAGTCAACCTGA	30,802,357
BIEC2-692880	GAGCATGGGATGTTGTAGTT	GGGATGTCACTGGACCTCTCT	30,817,514
BIEC2-692882	AGTGCCTAGAACAGTGCCAGA	TTTTACTGGGGCTCGGAGAT	30,820,966
BIEC566827A*	CGAACTCTGGCTCTCATTC	TGCAGCTGTGGACCAGACA	30,976,629
BIEC566819	AGCAGGGCAATAGAAGAGGAA	AATGCACCCATTTTGAGTGTG	31,010,210
BIEC566818	GAGAAGGGCTGCAAACTACTCA	TGCTTTGGGTGGAACCTTT	31,027,232
BIEC566817	GAGAAGGGCTGCAAACTACTCA	GTGCCTTGGGTGGAACCTTT	31,027,256
BIEC2-692941	GCCAGGGAGCATAGGTGTATT	CTGAAAAGGCAGGTCCAAAC	31,068,159
BIEC566795	ATGCAAAGAACCACCACAGAA	GGTGAGAGATGGCAGCAATG	31,068,177
BIEC566974	ATGCAAAGAACCACCACAGAA	GGTGAGAGATGGCAGCAATG	31,068,196
BIEC2-692944	GCCAGGGAGCATAGGTGTATT	CTGAAAAGGCAGGTCCAAAC	31,068,215
BIEC566792	ATGCAAAGAACCACCACAGAA	GGTGAGAGATGGCAGCAATG	31,068,320
BIEC566748	AAATCCACCCATTTTCTCCAC	GAGAACACTACCTTTAAATCCCAAT	31,232,127
BIEC566735	GCTTAGGATCCATGTGCT	TCCATCTTCAATATTATTAATCCCTAA	31,339,803
BIEC566734A*	GCTTAGGATCCATGTGCT	TCCATCTTCAATATTATTAATCCCTAA	31,339,899
BIEC566734	GCTTAGGATCCATGTGCT	TCCATCTTCAATATTATTAATCCCTAA	31,339,957
BIEC566707	ACGGGAAATAAATTGGCTCTGT	CCAGTGGCATAGTGGTTGAGT	31,551,084
BIEC566706A*	ACGGGAAATAAATTGGCTCTGT	CCAGTGGCATAGTGGTTGAGT	31,551,113
BIEC566706	ACGGGAAATAAATTGGCTCTGT	CCAGTGGCATAGTGGTTGAGT	31,551,245
BIEC566704	ACGGGAAATAAATTGGCTCTGT	CCAGTGGCATAGTGGTTGAGT	31,551,406
BIEC566703	ACGGGAAATAAATTGGCTCTGT	CCAGTGGCATAGTGGTTGAGT	31,551,409
BIEC566678	TCCTTCGTCTGTGTCTGACCT	GGAAGAGAGAAAAGCAAAATTCATC	31,641,072
BIEC566660	TCGGTCAAGAGATTGCTGACT	CAAATATCTTGGATAGGAAACACTTT	31,745,919
BIEC566659	TCGGTCAAGAGATTGCTGACT	CAAATATCTTGGATAGGAAACACTTT	31,745,956
BIEC566644A*	CCAACAGCCTGAAAGAAACTG	AACGTTATTTTGTACTCCTTGG	31,802,944
BIEC566644	CCAACAGCCTGAAAGAAACTG	AACGTTATTTTGTACTCCTTGG	31,803,062
BIEC566636	CCCAAAGTCCCAGCTAAATTC	TGATGATGATGGATAGCTTTGAA	31,903,435
BIEC566635	CCCAAAGTCCCAGCTAAATTC	TGATGATGATGGATAGCTTTGAA	31,903,459
BIEC2-693113	TTAGTGAACGTTGGGATCCAG	TGGCTCTTTTCTCTGTTGAG	32,044,379
BIEC566599	GCTTCGGTAGCAAAGAATCAC	ACCTGCAAGCCAAGAACAC	32,206,309
BIEC566598A*	GCTTCGGTAGCAAAGAATCAC	ACCTGCAAGCCAAGAACAC	32,206,328
BIEC566598	GCTTCGGTAGCAAAGAATCAC	ACCTGCAAGCCAAGAACAC	32,206,347

BIEC566597	GCTTCGGTAGCAAAGAATCAC	ACCTGCAAGCCAAAGAACAC	32,206,352
Primers used for candidate mutation screening (11 novel variants identified from 454 re-sequencing)			
Difference 1	TTGGGTCGAGTCTCTTTCTCA	CTTTTGTCTCACGCATTTTGC	30,403,703
Difference 2	TGGGTGATCCCATGACTTTA	GAAGTGGGGAGGACCTGTAG	30,407,557
Difference 3	AAATATCGGGGAAGATTGCTG	CTTCAGGAGCATGAACTCTGG	30,420,188
Difference 4	GTGGGATTCCAGAAGGAGAG	TTCCCCTTTCCCTCTACAC	30,524,595
Difference 5	TTCACTTCGGTTATGCCTTTG	AAGCCGGGTACCTGAGAGTTA	30,529,341
Difference 6	GCTGGTTTTCTGGTGATGGTA	CTCTTAAGCCAAAACCCACA	30,529,406
Difference 7	GATTCGGTGCTCTTACCATGA	TGCACCATCTCTCATCTTCT	30,632,474
Difference 8	TTTTAGGAATAAGGGCCTTCG	CCCATGGCGTAGTGTTAAGT	30,649,227
Difference 9	CTCATGATTGTGGGGAGGATA	ATCAGGTTGGTCACATTCTGG	30,660,224
Difference 10	CCCCTAAGCACACACTATCCA	TAACATCTGCCACCAATCCTC	30,696,518
Difference 11	AGGCCAAGTCAACAAAAGGTT	TTGGATCAAAGAGGTGGTGTC	30,697,396

All SNPs labeled with a BIEC number were identified from those SNPs on chromosome 26 which are publically available online from the Broad Institute (<http://www.broadinstitute.org/mammals/horse/snp>). Those BIEC SNPs which have an asterisk are novel SNPs which were identified in the same amplicon as the SNP which was identified from the Broad Insititute e.g. when BIEC566644 was sequenced, a second SNP (BIEC566644A) which was novel was also identified.

Primer table for those primers used to manually sequence across the sequence gaps in the 454 re-sequencing data. Start and finish refer to the coordinates of the gap within the sequencing data.

Start	Finish	Forward Primer	Reverse Primer
30,373,041	30,373,111	GCAACGTTTCAGGTTGATGAT	GGGAAAGGAGAATGGAGAGTG
30,373,297	30,373,306	GCAACGTTTCAGGTTGATGAT	GGGAAAGGAGAATGGAGAGTG
30,375,044	30,375,071	ACAGTTAGCGTGTGTCTCCT	AGGCCACATACTGCATGATTC
30,375,217	30,375,416	TTACACCCCAACCCTTAACAA	ATGCAGGGAACAAAGTCTGAA
30,376,177	30,376,441	CAGGCCCTTACAGTGTAGCTG	GCAATGTCAGCTCTCCTCACT
30,376,641	30,376,653	AGCTGACATTGCCACAGAAAT	CAATCCTTCATGCCTTCAAGA
30,377,778	30,378,016	GCAGGGCAGACAGACACTAAC	GCCTGTCTGACCACCTTATT
30,380,237	30,380,565	TCAGTGTGAGATTGTTTGTATGC	AACACCCGTTTCATGACAGAAA
30,381,170	30,381,181	CTTTGGAACCTGGAGCTTTCT	CAGACTTGCAAAGAACAACAGA
30,381,897	30,382,121	TGTGTGTGTGAGGAAGACTGG	GCCATAAAAATGAGGCTCGATA
30,382,937	30,383,086	GAGTCAGGCAGCCCTTCTTAT	GACAAACAACGAAGGACCAAG
30,386,354	30,386,504	GCCTTCAGTATCAGTACGTTGC	TTGGCACACTTCACCTTCTGTG
30,392,595	30,392,766	ATAAGCCTTCTCCGAAGAAA	GGGGCGCTTCTAAAATTAAC
30,393,163	30,393,326	CGCTCTGTGGTTGAGTTCAT	CTGTCCTTTCATTCTCTGCTCA
30,397,486	30,397,502	GTGCCCTCTGTCTTCTTTCT	CCACCCCTAGGAAAACCATAA
30,397,763	30,397,791	GTGCCCTCTGTCTTCTTTCT	CCACCCCTAGGAAAACCATAA
30,398,293	30,398,546	GTAAACATCCCTACCCCTCCA	AGGCAGCCCAGTTCTAGAGTC
30,403,589	30,403,831	TTGGGTCGAGTCTCTTTCTCA	GAAAGGCATTTCTTAGCCAGA
30,405,584	30,405,851	GGGAGAAACACATGAAAAAGC	TTCAGAAGTCTTCCCTGAGCA
30,405,952	30,405,999	TCTGAGCTAACTGCTGCCACT	TCTAATGGAAGCACCTGGAGA
30,406,082	30,406,150	TCTGAGCTAACTGCTGCCACT	TCTAATGGAAGCACCTGGAGA
30,407,176	30,407,866	ACCTGTTTGCTGTCCATGTTT	TCTAGAGGCTTCATGGCTGTC
30,410,972	30,410,984	TATGGGCCTTAGGACTTTTGG	GCAGACATGCTGCATAGACAT
30,416,705	30,416,876	CTCCACTGAAAATCCAAATCAG	CCAGGAGGTCGGATTTATTCT
30,423,026	30,423,038	ATTCAGGTTGCTGGAGGATTT	CCTACGGCATTCAAATGAGAG
30,432,886	30,432,956	ATGCAGCATTGCTCTGATTCT	GAAGGCAGTAGGTGATCATGC
30,433,412	30,433,431	ACACAGGGTAGTGGGTCCTTT	CACATGCAGCATTCTTGATA
30,439,260	30,439,280	TATCCTGGGCCATTAAGAGAGG	CACAAATGCCTTTTCTCTCCA
30,442,054	30,442,085	TGCAGTCACGCATCACTAAC	TCAACGAGGGACCACATAGAC
30,444,014	30,444,031	TATGGGCTGAACTGTGTTTCC	CCTTGCTGGAAGGTGATGTA
30,444,208	30,444,211	TATGGGCTGAACTGTGTTTCC	CCTTGCTGGAAGGTGATGTA
30,450,663	30,450,706	GAGTGTGCCTCCTGTATCAGC	GCTGCCATCAGAAGCTAAGAG
30,454,310	30,454,835	AGGAGGCATGGGGGTTAGTTA	CCAAGGAAAGGGGTTGGACT
30,456,334	30,456,356	CTGTGGAGAGACAGCAGAAG	CATCCTGTTTTACCTCCAA
30,463,225	30,463,241	GCTGCCTGGGTTTGAATCT	CTGCTTAGTATGAACCAAACACTTC
30,463,609	30,463,796	TGATTTGGCTTGCCAGATAG	TCCCACGTTGTACTCTGCTTT
30,464,074	30,464,086	TGGGGAGAAAAGCAGAGTACA	ATCACCCATAATCTCCCCATC
30,466,711	30,466,746	TCCACCATTTTCTTCTGTGCG	TGGTTTAACTCTCCACCTCA
30,466,941	30,466,956	TCCACCATTTTCTTCTGTGCG	TGGTTTAACTCTCCACCTCA
30,470,590	30,470,631	CATGCACTCAAACACAAGCAT	CAAACCAGACTGGGTGTCTGT
30,472,182	30,472,406	GAGGGATGGAGGGCAAAT	CATTTGATGCCTTTCCATT

Start	Finish	Forward Primer	Reverse Primer
30,474,192	30,474,465	GGGCTCTTAACTCCTGTGGTT	CCAATGTTACTTCACAC AACAGATG
30,474,801	30,478,171	TCGGTGGAGAGTCTACAGTG	TGAGAAAAGAGACAAAGATGTGGAG
30,478,239	30,478,670	AAGTGACCCATTGCTAGTTCAA	AATTCAGGTAAGCGAGACCA
30,478,747	30,479,026	ATACTGACCCCTCCTGGTCTCG	GGGTACACCAAAACACATAGGA
30,479,102	30,479,591	CTGGTTCTCCTGCATTTCCTT	CAGGCTTTCAGCAGATGTCTC
30,479,689	30,479,736	TTCCATTCTTCTAGGCTGT	CTGGGAGGTGATACAAAACCA
30,482,768	30,482,966	GGAGAGAGGAATGTGGGAGAG	AGTGGCCTAGACAAAACCTGT
30,491,503	30,491,526	TAATATGCATTGAGCCCTTCC	GCAACCGAAACATGTAACCAT
30,495,851	30,496,003	GATGCAGGCCTCTCTGTGTAG	TCTTTGCCCCTGCCGTCT
30,500,713	30,500,724	AGCCCCGAAATCTTGGTTTCTA	GCGGTAAGAAGCTTTCCTGTT
30,502,380	30,502,446	ACCACCTTTGAGCAGGATTTT	GTTCAAGTTCGCACCTCTCC
30,504,405	30,504,426	CCCTTAGGGTGCAGTTTCTC	ACATTGGGAGACACACAAAA
30,510,919	30,510,921	TGGTGAGGAAGAATTTGCTGA	GCATGGATTGAGGAAACTGA
30,511,095	30,511,193	TGGTGAGGAAGAATTTGCTGA	GCATGGATTGAGGAAACTGA
30,512,609	30,512,711	AAACCTGTCTTCCACCTACC	GGAATTTCTTTTGGCCTGA
30,513,756	30,513,816	GTTTACCCCTGCGGTATCTGT	TTTGTGCAACCATCACCCTA
30,513,990	30,514,316	GAGGCAAACCCATAGAAATGA	TTGTACCCTCAAGGGCTTT
30,516,014	30,516,026	GATATTTACAGGGCCCTTACG	TTTATCCAATCCCAAGAACC
30,524,497	30,524,698	TCAAGAAGCCCAATAAGTTCTG	CTCCTTCTATTGGATTGTTTGG
30,525,109	30,525,311	TCACTACATGTGCTAAAAGGAACTG	GGACTTGTCAATCTTTCCTGA
30,526,386	30,526,666	CTGATACTGGAACCCACAAA	GCATTGTATCCTGCACGTTTT
30,529,333	30,529,471	GCTGGTTTTCTGGTGATGGTA	TGCCCTAACTACACACAATTTCG
30,529,671	30,529,689	GGGCAAGTATTGTTTCAGTGC	TAGGGTCATAGGAGGGGAAGT
30,534,060	30,534,271	AACACTTGACCTTGGCATTG	GTATGCTTTTTGGTGAGGAAGA
30,534,343	30,534,386	AGAGGAAGTTTGGAACAGGT	GCAAAAAGCAAGCTCACGATA
30,534,465	30,534,486	AGAGGAAGTTTGGAACAGGT	GCAAAAAGCAAGCTCACGATA
30,535,776	30,535,790	TCACTCGACGCATTCATCATA	GTTTGACAGGCCGTAAGTCAG
30,536,394	30,536,571	CAATGGGAAGCTATTGAACCA	TGGGTTAATCGAGTCTGGA
30,537,042	30,537,201	AATCAGACTCTGCACCAGACG	GGGTAGGTTTGGAAAGCGTTTA
30,540,905	30,540,911	AGTTTCCTATCGCCCCTGTAA	GTGTGGTAGGCTGAGAAATGG
30,541,065	30,541,076	AGTTTCCTATCGCCCCTGTAA	GTGTGGTAGGCTGAGAAATGG
30,541,225	30,541,231	AGTTTCCTATCGCCCCTGTAA	GTGTGGTAGGCTGAGAAATGG
30,542,856	30,543,111	GTGACCAGTCATGCCAGTTTG	AGCCACTCTCTCCAAGTCC
30,546,796	30,546,846	CTTCCACTGTGGCTTTTGAGA	ATGCTCATCCACAGAATCCAG
30,549,926	30,549,951	GCACAGAGAGCATGGAAGAG	CCAGTCACAAAAGGGCAAATA
30,550,048	30,550,050	GCACAGAGAGCATGGAAGAG	CCAGTCACAAAAGGGCAAATA
30,550,129	30,550,300	AGCCACTGGTAACCACCTTT	ACAGTGTGGTGGTTCTCCTAAA
30,550,404	30,550,466	TCATCTGCTGATGGACATTTG	CCAATGCCTACAAACCAAGA
30,551,274	30,551,311	CAGGGCCAAACCTAGTATTCC	CCCTCCCTACTCCTTACATC
30,553,113	30,553,191	TTTCATACCCCAAACAGAAG	ATTGAAACCTTACACGTTGC
30,556,211	30,556,281	TTGAAGTCCGAGATCAAGGTG	CAGTCGAGAGGCCCTAAGTCT
30,558,345	30,558,401	AGAGAAGCACAGAGGGAGGAC	TCTTCGCCCTGAGAGTATCAA
30,559,028	30,559,036	TTTCTAGCTTCTGCACCAGGA	TGAAGCCATAGCATCACAAACA
30,559,181	30,559,196	TTTCTAGCTTCTGCACCAGGA	TGAAGCCATAGCATCACAAACA
30,561,692	30,561,781	AGAAGTGAAAGGTGGAGAGG	AGCCTTTCAGTTTGGAAAAGG
30,561,874	30,561,906	ATGACTGCTGAGGAGTGCTGT	GAAGAATTCGCACTGTTGTCC
30,563,567	30,563,801	TTCTGGGTCAATCCAGTATGC	GGACAATGACTGGAGGTAGCA

Start	Finish	Forward Primer	Reverse Primer
30,566,316	30,566,326	TGCACCATGACAGGAATTGTA	CTCTTGGATTCTGGCCTACC
30,566,483	30,566,531	TGCACCATGACAGGAATTGTA	CTCTTGGATTCTGGCCTACC
30,575,067	30,575,096	TGGGAAGACAAGTCATGTTGA	AGAGGAAAGTGGGTGTGGAGT
30,576,475	30,576,526	CCAACTGGGGACATACATCAC	AGCTGTCATTGCTGTTCTGCT
30,577,676	30,577,691	TGACATCAGTGAGGAACACATT	CCTTCATGTCTTCCGTAGTTTG
30,577,883	30,577,899	TGACATCAGTGAGGAACACATT	CCTTCATGTCTTCCGTAGTTTG
30,577,985	30,578,106	AAGCAACCAAGATGCTCTTCA	CATGGCCCTAAAAATCCTCTC
30,578,220	30,578,296	AAAGGCAAAACTGTGGAGACA	CCGATGTGTTGGTGTAAAGGT
30,578,522	30,578,532	GCTGCCATTAACCTTAACACC	ACAATTCCATCCAGAGCTCAC
30,578,740	30,578,896	CCCTTTGTGAACAGCCAGAT	AAGTCCCACCCACTTACATC
30,580,504	30,580,654	CCATGTCCTCTTGCCACTCTA	AGTTTGCCTAAGCCAAGGACT
30,581,139	30,581,151	CCACTGCCTGTTTTGTCAAT	ACGATCAGGGTCTTACACAGC
30,585,711	30,585,730	CCCTGCCACCACAAAGTTTA	GAGTCCTCGCTCAGAAGACAA
30,589,599	30,589,611	AGAATTGTGCACCAACCAAGA	TTCCATTCTAGCACCATCTGC
30,591,799	30,591,870	GAGGCACAGATCAAAGTGAGG	TCTCTTACCCTTTGGGGAGAA
30,592,014	30,592,021	GAGGCACAGATCAAAGTGAGG	TCTCTTACCCTTTGGGGAGAA
30,592,723	30,592,761	CCTGAGAGCGACACAGAAAAC	CTTGGCTCACCAACATCAAAAT
30,598,190	30,598,371	GATCAGAGACTTCAGTGACCTCAT	CCAATGTCTGTGCTTCCTCAG
30,598,616	30,598,670	TCCCACACAAGGAAACAAATC	CAGCAGCTGGGATTTTGATAG
30,598,745	30,598,786	TCCCACACAAGGAAACAAATC	CAGCAGCTGGGATTTTGATAG
30,599,215	30,599,301	GAAAGACACGTAGACCATGGAA	GCCCCCTTTTCATTTTCTTG
30,599,466	30,599,976	ATCTTCATGACCTTGATTGG	AGGGGATCAGTTTCTCCACAT
30,600,309	30,600,376	GGGATGGCTATCCTCAAAGAC	TCAGTTGTGCAACCATTACCA
30,600,512	30,600,531	AGGCCACAAGTTGTATGATGC	TCAGTTGTGCAACCATTACCA
30,603,052	30,603,246	GCATTTGAAAAGTGCAGATGAG	CGGTATCATTAACTGCATTAGCC
30,603,356	30,603,436	TGAGGAAATGGGAGATGTCAG	TTTCCCAGCTTACTGAGGT
30,604,159	30,604,396	CTCGAATGGCAAAGTAACGAG	AAGCCAGTGGAGAAGGAAAAA
30,604,954	30,605,119	GGGCGAAGCTCTTTCAGTATT	CAGAAGAGCTCAGTTATTCTCAG
30,605,768	30,605,920	GGCATCAAAAAGATAAGAGGAG	GGGATTTTTGCTTCTGGATT
30,607,385	30,607,551	TGGTGAGGACGTAAAGGGAAT	AGTTTGAGGCTGTTGGGAACT
30,608,155	30,608,322	AGACTTCCGGGATGTGCTAAT	TAGGCTTGGTCTCCTTCCAT
30,608,544	30,608,566	TAGGGAGAGGACTGGGTGCT	CCAGGTGAAGTGCTGTAGAC
30,610,145	30,610,165	GATTGGTGGGCATGAGTGATA	CGGATTTACCTGTCCTTCTC
30,611,929	30,612,081	CATTTCAGTACCACCAGTT	CTGGGCATCTCAGTTCACACT
30,616,345	30,616,356	GGGGTCCAGTTTAACTCCAAA	TGTGTGCTACCACCACATCAT
30,617,067	30,617,101	CATCAGGTGTTGGAGGTGTTT	GCAGGTGAATGGATGAAGAAA
30,617,331	30,617,336	CCATTCACCTGCTGAAGGATA	TCCAGAACACTGAAAGCATCA
30,617,498	30,618,106	CAAAGCGGCTGTACCATTCTA	CAGCCCCGTCTCCATAATTT
30,618,194	30,618,496	TCTTCTTCCATGGATTGTGC	TGTGCCACCTATTCATCCTTC
30,618,607	30,618,690	GCTTTGGGGAGAAGAAGAAGA	CTGGCCCTCACTTGAGATTTT
30,618,818	30,618,871	GCTTTGGGGAGAAGAAGAAGA	CTGGCCCTCACTTGAGATTTT
30,619,000	30,619,171	TCCTCTGCACATACAAAACCTTC	TGGCATAGTTCAGTTGGAGGA
30,622,625	30,622,796	CTGCTTCACTTCCACCTCAAG	GCTGCTGATGGGAATAGAAAA
30,623,001	30,623,455	TAGATGGCCAGGTCCTATGGT	GAAACAAGTCAGAGACCAGGAGA
30,624,157	30,624,181	AGATTGCTGTGTGCAAATCCT	GCAGGAAATTATATGAGCCAAC
30,626,011	30,626,064	TCCACAAAGGAAACATCCATC	TCCCTTCGCTAATCTACTCC
30,626,140	30,626,180	TCCACAAAGGAAACATCCATC	TCCCTTCGCTAATCTACTCC

Start	Finish	Forward Primer	Reverse Primer
30,630,877	30,631,071	TGAGCACCTTTTCACATACCTG	CAACCTACAAAATGGGAGCAA
30,630,232	30,630,343	TCTGAGTAATATCTCCCCATTTC	ATCTGCACTCCCATGTTTCATC
30,630,539	30,630,806	GATGAACATGGGAGTGCAGAT	CTCAGGGCCAATCTTCCTAAC
30,631,146	30,631,186	CGGAGCATTCAAACCTTACCA	ACTTCTTGACATTTGCCTTGG
30,631,291	30,631,466	TTGCTCCCATTTTGTAGGTTG*	GGAACAGAATAGAGAGGCCAGA*
30,631,576	30,632,021	GGATATCCAGTTTTCCCAACTA*	CCATGTTTCATGGATGGGAAG*
30,632,096	30,632,116	TTCCCTCAAATGCTTTGG	AAGCTTAACCAAGGAGGTGAAA
30,632,309	30,632,545	TGTACAGATATTTACCTCCTTGG	TGCCTTGAGATTGATTCTAACCC
30,635,027	30,635,716	GTGGGATTATCCAGCAGATGA*	CGTCGGAAAGAGAGAAGTCAA*
30,640,299	30,640,305	ACTGCTTGTGGAAAGTCAACG	AGTGTGCCAACTCCTGTCTA
30,640,382	30,640,466	ACTGCTTGTGGAAAGTCAACG	AGTGTGCCAACTCCTGTCTA
30,644,677	30,644,751	GGAGATGGCAATGTGAAGTGT	GCCATAACAGACGGGAGTGTA
30,644,847	30,644,941	GGAGATGGCAATGTGAAGTGT	GCCATAACAGACGGGAGTGTA
30,646,697	30,646,931	TCTGCCCTGTTTGTAGTTTGT	TGGGTCGCATATTTACCTAACCA
30,649,203	30,649,349	CCTTCGTCCTAACCAACAT	TCAAGGAAACCAAGGAACAAC
30,651,753	30,652,166	CAGAGGCGTGTGGTACATT	CCCTTCCACCTGAGATAAAAA
30,652,781	30,652,784	GGACAGACACATGCAAATCCT	TTCATCCACACACAGACATCC
30,653,205	30,653,209	GGACAGACACATGCAAATCCT	TTCATCCACACACAGACATCC
30,655,951	30,656,166	CAGCTTGAGCTAGTGTGACC	GCTGGGATCAGTAACAGTGCT
30,663,739	30,663,750	GTGAATTGCTGATTGCAGCTT	TCTCCACCTTTGAAGACAAGTG
30,674,188	30,674,356	GCCACCACCCACAAGAAGTAT	AGAATGGAAGCGGTCTGATTT
30,674,995	30,675,051	GCCTGACTCAGATCCACAAG	TTCCCTAACTAAGCCCCTTCA
30,678,122	30,678,296	CCCCTCCCATATTACTGCCTA	CCCTCAGCCACTTCTAGGAAC
30,681,409	30,681,586	TGCACAAGGCTCTGTTCTCT	ATGTAGTGCCAGAATGCTGCT
30,688,396	30,688,586	TGGGTTGCTCTCACGTAGTTC	CCAGACAGTGATTTATCCAGA
30,696,359	30,696,366	TGCTGTAAAACCTTATGCTGTG	GCCTTCATCTCCAAGATTGTG
30,696,449	30,696,663	TGCTGTAAAACCTTATGCTGTG*	GCCTTCATCTCCAAGATTGTG*
30,697,087	30,697,109	CTCATTGAAGGCAGTGAGACC	TGTTGACTTGGCCTCCTTATG
30,697,387	30,697,651	TCAAGGCTTTTAAGGACAGGA*	CCTCCTGAGTGCTTCTTCCA*
30,697,761	30,697,765	ATTTGCTGCAACTGAGGAAGA	GCAAATGCTTCTGCTAAATCC
30,697,847	30,697,996	ATTTGCTGCAACTGAGGAAGA*	GCAAATGCTTCTGCTAAATCC*
30,698,262	30,698,315	AAAGGAACTGCAGGAGAAGG	GGACACTCCTGTGGACTGTGT
30,698,431	30,698,476	AAAGGAACTGCAGGAGAAGG	GGACACTCCTGTGGACTGTGT
30,698,954	30,698,981	ACACAGTCCACAGGAGTGTCC	GGATTGCAACCATCCAACCTTA
30,699,785	30,699,815	TGTACCCCATCCCTAGAAAC	CTGAAGAGCATTTCACCGAAG
30,699,895	30,699,936	TGTACCCCATCCCTAGAAAC	CTGAAGAGCATTTCACCGAAG
30,700,157	30,700,192	AGCGGGGATATGTTGTCTTT	TGAACTCAAAAATGCAGCACAG
30,700,279	30,700,285	AGCGGGGATATGTTGTCTTT	TGAACTCAAAAATGCAGCACAG
30,701,343	30,701,391	CTTTCAGCAACGTTTTGTGGT	TCGCAGCAAAGTGGGATAGA
30,701,788	30,701,851	CTTGGGCTAAACCCTACTTGG	TCCCGCAAAGAAAACCTACAGA
30,705,963	30,706,050	CCAAAGAGTAGCTGGCACAGA	ACAACGCATCAGTTGTTGCTA
30,713,379	30,714,220	AACTAAACAGATTCATGGTCTGGAG*	TCAAGGGGTCAATCCACATAA*
30,715,460	30,716,251	GTGATGCTACAGAGGGCTCAC*	AGGGAGCTTACATTCACCTG*
30,718,181	30,718,241	AGGCCAGGCTTTTGTTAGAG	TTCTGGAAAGTCTCCCTGGAT
30,721,738	30,721,855	AAAAATTAGCCACAGGGTTCC	GGCTATCATGAGAATCAATGTAGG
30,723,017	30,723,061	GGCCTGATTGTTATGGACTGA	ATCCAAGATCAGAGTGCCAGA
30,725,181	30,725,265	TTCTTTTCCCTTGTGCTCT	ACTTTTACTCAGGCGCTCGAC

Start	Finish	Forward Primer	Reverse Primer
30,725,395	30,725,471	TTCCTTTTCCTTTGTGCTCT	ACTTTTACTCAGGCGCTCGAC
30,730,683	30,730,700	CCAGACGGTTTCTACTAGCAAAT	CCATTCACCCACTAAAGGACA
30,731,105	30,731,136	CCAGACGGTTTCTACTAGCAAAT	CCATTCACCCACTAAAGGACA
30,740,192	30,740,284	GAACACAAACAAAGGGGGATT	TAGCGGTTTGCTGACAGTCTT
30,740,519	30,740,601	TTGGACACAGAGACAGACGTG	GTGCTGAGAAATCGGTATTGG
30,743,327	30,743,420	CAGAAATGGCATTCTTCCAG	AGCCAAGGGTGTAGTTCCAGT
30,744,461	30,744,671	GCGCACACAAATATCTGGTTA	AATTCCTTCTTAACCCCAAGG

* Primers which failed to amplify the amplicon and so this gap in the sequencing remains.

APPENDIX 12**ABBREVIATIONS, BUFFERS AND REAGENTS****Abbreviations:**

FPS	-	Fell Pony Syndrome
FIS	-	Foal Immunodeficiency Syndrome
SNP	-	Single nucleotide polymorphism
ECA26	-	Equine chromosome 26
LD	-	Linkage disequilibrium
GWAS	-	Genome-wide association study
IBD	-	Identical-by-descent
PCR	-	Polymerase chain reaction
gDNA	-	Genomic DNA

Buffers and Reagents:

Nucleon A Reagent - 10 mM Tris-HCl, 320 mM sucrose, 5 mM MgCl₂, 1% (v/v) Triton X-100; pH to 8.0 with 40 % (w/v) NaOH

Nucleon B Reagent - 6.34 mM Trizma Hydrochloride, 2.70 mM EDTA, 17 mM NaCl, 800 mL ddH₂O; pH to 8.0 with 2M NaOH and then add 10g SDS

TE Buffer – TE, 10mM Tris-HCL (pH 7.5), 0.1mM EDTA

TAE Buffer (10X) - 400mM Tris-acetate, 10mM EDTA (pH 8)

Bibliography

-
- ABBOTT, J. C., AANENSEN, D. M. & BENTLEY, S. D. (2007) WebACT: an online genome comparison suite. *Methods Mol Biol*, 395, 57-74.
- AL-MUHSEN, S. & CASANOVA, J. L. (2008) The genetic heterogeneity of mendelian susceptibility to mycobacterial diseases. *J Allergy Clin Immunol*, 122, 1043-51; quiz 1052-3.
- ANDERS, S. & HUBER, W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, 11, R106.
- ANDERSSON, L. S., JURAS, R., RAMSEY, D. T., EASON-BUTLER, J., EWART, S., COTHRAN, G. & LINDGREN, G. (2008) Equine Multiple Congenital Ocular Anomalies maps to a 4.9 megabase interval on horse chromosome 6. *BMC Genet*, 9, 88.
- AWANO, T., JOHNSON, G. S., WADE, C. M., KATZ, M. L., JOHNSON, G. C., TAYLOR, J. F., PERLOSKI, M., BIAGI, T., BARANOWSKA, I., LONG, S., MARCH, P. A., OLBY, N. J., SHELTON, G. D., KHAN, S., O'BRIEN, D. P., LINDBLAD-TOH, K. & COATES, J. R. (2009) Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proc Natl Acad Sci USA*, 106, 2794-9.
- BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21, 263-5.
- BAU, S., SCHRACKE, N., KRANZLE, M., WU, H., STAHLER, P. F., HOHEISEL, J. D., BEIER, M. & SUMMERER, D. (2009) Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal Bioanal Chem*, 393, 171-5.
- BELL, S. C., SAVIDGE, C., TAYLOR, P., KNOTTENBELT, D. C. & CARTER, S. D. (2001) An immunodeficiency in Fell ponies: a preliminary study into cellular responses. *Equine Vet J*, 33, 687-92.
- BELLONE, R. R., FORSYTH, G., LEEB, T., ARCHER, S., SIGURDSSON, S., IMSLAND, F., MAUCELI, E., ENGENSTEINER, M., BAILEY, E., SANDMEYER, L., GRAHN, B., LINDBLAD-TOH, K. & WADE, C. M. (2010) Fine-mapping and mutation analysis of TRPM1: a candidate gene for leopard complex (LP) spotting and congenital stationary night blindness in horses. *Brief Funct Genomics*, 9, 193-207.
-

- BENJAMINI, Y. & HOCHBERG, Y. (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of Royal Statistical Society*, 57, 289-300.
- BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA, E. C. M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53-9.
- BERNOCO, D. & BAILEY, E. (1998) Frequency of the SCID gene among Arabian horses in the USA. *Anim Genet*, 29, 41-2.
- BERRY, G. T., WU, S., BUCCAFUSCA, R., REN, J., GONZALES, L. W., BALLARD, P. L., GOLDEN, J. A., STEVENS, M. J. & GREER, J. J. (2003) Loss of murine Na⁺/myo-inositol cotransporter leads to brain myo-inositol depletion and central apnea. *J Biol Chem*, 278, 18297-302.
- BRANDL, C. J. & DEBER, C. M. (1986) Hypothesis about the function of membrane-buried proline residues in transport proteins. *Proc Natl Acad Sci USA*, 83, 917-21.
- BRENNER, S., JOHNSON, M., BRIDGHAM, J., GOLDA, G., LLOYD, D. H., JOHNSON, D., LUO, S., MCCURDY, S., FOY, M., EWAN, M., ROTH, R., GEORGE, D., ELETR, S., ALBRECHT, G., VERMAAS, E., WILLIAMS, S. R., MOON, K., BURCHAM, T., PALLAS, M., DUBRIDGE, R. B.,

-
- KIRCHNER, J., FEARON, K., MAO, J. & CORCORAN, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18, 630-4.
- BRIGHT, L. A., BURGESS, S. C., CHOWDHARY, B., SWIDERSKI, C. E. & MCCARTHY, F. M. (2009) Structural and functional-annotation of an equine whole genome oligoarray. *BMC Bioinformatics*, 10 Suppl 11, S8.
- BROOKS, S. A., GABRESKI, N., MILLER, D., BRISBIN, A., BROWN, H. E., STREETER, C., MEZEY, J., COOK, D. & ANTCZAK, D. F. (2010) Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for Lavender Foal Syndrome. *PLoS Genet*, 6, e1000909.
- BROWNE, S., SULLIVAN LS, KOBOLDT DC, DING L, FULTON R, ABBOTT RM, SODERGREN EJ, BIRCH DG, WHEATON DH, HECKENLIVELY JR, LIU Q, PIERCE EA, WEINSTOCK GM AND DAIGER SP (2010) Identification of Disease-Causing Mutations in Autodominant Dominant Retinitis Pigmentosa (adRP) Using Next-Generation DNA Sequencing. *Investigative Ophthalmology and Visual Science*, 10, 6180
- BRUNBERG, E., ANDERSSON, L., COTHRAN, G., SANDBERG, K., MIKKO, S. & LINDGREN, G. (2006) A missense mutation in PMEL17 is associated with the Silver coat color in the horse. *BMC Genet*, 7, 46.
- BUCCAFUSCA, R., VENDITTI, C. P., KENYON, L. C., JOHANSON, R. A., VAN BOCKSTAELE, E., REN, J., PAGLIARDINI, S., MINARCIK, J., GOLDEN, J. A., COADY, M. J., GREER, J. J. & BERRY, G. T. (2008) Characterization of the null murine sodium/myo-inositol cotransporter 1 (Smit1 or Slc5a3) phenotype: myo-inositol rescue is independent of expression of its cognate mitochondrial ribosomal protein subunit 6 (Mrps6) gene and of phosphatidylinositol levels in neonatal brain. *Mol Genet Metab*, 95, 81-95.
- BUCKLEY, R. H. (1986) Humoral immunodeficiency. *Clin Immunol Immunopathol*, 40, 13-24.
- BURG, M. B., FERRARIS, J. D. & DMITRIEVA, N. I. (2007) Cellular response to hyperosmotic stresses. *Physiol Rev*, 87, 1441-74.
- BURROUGHS, L., WOOLFREY, A. & SHIMAMURA, A. (2009) Shwachman Diamond Syndrome - a review of the clinical presentation, molecular pathogenesis, diagnosis, and treatment. *Hematol Oncol Clin North Am.*, 23, 233-248.
-

-
- CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M. A., BARRELL, B. G. & PARKHILL, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, 21, 3422-3.
- 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.
- CHAU, J. F., LEE, M. K., LAW, J. W., CHUNG, S. K. & CHUNG, S. S. (2005) Sodium/myo-inositol cotransporter-1 is essential for the development and function of the peripheral nerves. *FASEB J*, 19, 1887-9.
- CHURCHILL, G. A. & DOERGE, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, 138, 963-71.
- CLAYTON, D. G., WALKER, N. M., SMYTH, D. J., PASK, R., COOPER, J. D., MAIER, L. M., SMINK, L. J., LAM, A. C., OVINGTON, N. R., STEVENS, H. E., NUTLAND, S., HOWSON, J. M., FAHAM, M., MOORHEAD, M., JONES, H. B., FALKOWSKI, M., HARDENBOL, P., WILLIS, T. D. & TODD, J. A. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37, 1243-6.
- COOK, D., BROOKS, S., BELLONE, R. & BAILEY, E. (2008) Missense mutation in exon 2 of SLC36A1 responsible for champagne dilution in horses. *PLoS Genet*, 4, e1000195.
- CORDELL, H. J. & CLAYTON, D. G. (2005) Genetic association studies. *Lancet*, 366, 1121-31.
- CRISMAN, M. V. & SCARRATT, W. K. (2008) Immunodeficiency disorders in horses. *Vet Clin North Am Equine Pract*, 24, 299-310, vi.
- D'ASCENZO, M., MEACHAM, C., KITZMAN, J., MIDDLE, C., KNIGHT, J., WINER, R., KUKRICAR, M., RICHMOND, T., ALBERT, T. J., CZECHANSKI, A., DONAHUE, L. R., AFFOURTIT, J., JEDDELOH, J. A. & REINHOLDT, L. (2009) Mutation discovery in the mouse using genetically guided array capture and resequencing. *Mamm Genome*, 20, 424-36.
-

-
- DAWN TEARE, M. & BARRETT, J. H. (2005) Genetic linkage studies. *Lancet*, 366, 1036-44.
- DE SCHRIJVER, J. M., DE LEENEER, K., LEFEVER, S., SABBE, N., PATTYN, F., VAN NIEUWERBURGH, F., COUCKE, P., DEFORCE, D., VANDESOMPELE, J., BEKAERT, S., HELLEMANS, J. & VAN CRIEKINGE, W. (2010) Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. *BMC Bioinformatics*, 11, 269.
- DE WEERD, N. A., SAMARAJIWA, S. A. & HERTZOG, P. J. (2007) Type I interferon receptors: biochemistry and biological functions. *J Biol Chem*, 282, 20053-7.
- DEIGHTON, R. F., KERR, L. E., SHORT, D. M., ALLERHAND, M., WHITTLE, I. R. & MCCULLOCH, J. (2010) Network generation enhances interpretation of proteomic data from induced apoptosis. *Proteomics*, 10, 1307-15.
- DIXON, J. B., SAVAGE, M., WATTRET, A., TAYLOR, P., ROSS, G., CARTER, S. D., KELLY, D. F., HAYWOOD, S., PHYTHIAN, C., MACINTYRE, A. R., BELL, S. C., KNOTTENBELT, D. C. & GREEN, J. R. (2000) Discriminant and multiple regression analysis of anemia and opportunistic infection in Fell pony foals. *Vet Clin Pathol*, 29, 84-86.
- DROEGE, M. & HILL, B. (2008) The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*, 136, 3-10.
- DROR, Y. (2005) SHwachman-Diamond Syndrome. *Pediatric Blood Cancer*, 45, 892-901.
- ELLEGREN, H., JOHANSSON, M., SANDBERG, K. & ANDERSSON, L. (1992) Cloning of highly polymorphic microsatellites in the horse. *Anim Genet*, 23, 133-42.
- ESCAMILLA, M. A., DEMILLE, M. C., BENAVIDES, E., ROCHE, E., ALMASY, L., PITTMAN, S., HAUSER, J., LEW, D. F., FREIMER, N. B. & WHITTLE, M. R. (2000) A minimalist approach to gene mapping: locating the gene for acheiropodia, by homozygosity analysis. *Am J Hum Genet*, 66, 1995-2000.
- EXCOFFIER, L., LAVAL, G. & SCHNEIDER, S. (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online*, 1, 47-50.
-

-
- FAHAM, S., WATANABE, A., BESSERER, G. M., CASCIO, D., SPECHT, A., HIRAYAMA, B. A., WRIGHT, E. M. & ABRAMSON, J. (2008) The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na⁺/sugar symport. *Science*, 321, 810-4.
- FARCOMENI, A. (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat Methods Med Res*, 17, 347-88.
- FONDON, J. W., 3RD & GARNER, H. R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, 101, 18058-63.
- FOX-CLIPSHAM, L., SWINBURNE, J. E., PAPOULA-PEREIRA, R. I., BLUNDEN, A. S., MALALANA, F., KNOTTENBELT, D. C. & CARTER, S. D. (2009) Immunodeficiency/anaemia syndrome in a Dales pony. *Vet Rec*, 165, 289-90.
- FOX, S., FILICHKIN, S. & MOCKLER, T. C. (2009) Applications of ultra-high-throughput sequencing. *Methods Mol Biol*, 553, 79-108.
- FULLWOOD, M. J., WEI, C. L., LIU, E. T. & RUAN, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*, 19, 521-32.
- GEOR, R. J., JACKSON, M. L., LEWIS, K. D. & FRETZ, P. B. (1990) Prekallikrein deficiency in a family of Belgian horses. *J Am Vet Med Assoc*, 197, 741-5.
- GLOCKER, E. O., KOTLARZ, D., BOZTUG, K., GERTZ, E. M., SCHAFFER, A. A., NOYAN, F., PERRO, M., DIESTELHORST, J., ALLROTH, A., MURUGAN, D., HATSCHER, N., PFEIFER, D., SYKORA, K. W., SAUER, M., KREIPE, H., LACHER, M., NUSTEDE, R., WOELLNER, C., BAUMANN, U., SALZER, U., KOLETZKO, S., SHAH, N., SEGAL, A. W., SAUERBREY, A., BUDERUS, S., SNAPPER, S. B., GRIMBACHER, B. & KLEIN, C. (2009) Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med*, 361, 2033-45.
- GO, W. Y., LIU, X., ROTI, M. A., LIU, F. & HO, S. N. (2004) NFAT5/TonEBP mutant mice define osmotic stress as a critical feature of the lymphoid microenvironment. *Proc Natl Acad Sci U S A*, 101, 10673-8.
- GOLDSTEIN, O., ZANGERL, B., PEARCE-KELLING, S., SIDJANIN, D. J., KIJAS, J. W., FELIX, J., ACLAND, G. M. & AGUIRRE, G. D. (2006)
-

-
- Linkage disequilibrium mapping in domestic dog breeds narrows the progressive rod-cone degeneration interval and identifies ancestral disease-transmitting chromosome. *Genomics*, 88, 541-50.
- GOMBOS, Z., DANIHEL, L., REPISKA, V., ACS, G. & FURTH, E. (2011) Expression of erythropoietin and its receptor increases in colonic neoplastic progression: the role of hypoxia in tumorigenesis. *Indian J Pathol Microbiol*, 54, 273-8.
- GONG, Q. M., KONG, X. F., YANG, Z. T., XU, J., WANG, L., LI, X. H., JIN, G. D., GAO, J., ZHANG, D. H., JIANG, J. H., LU, Z. M. & ZHANG, X. X. (2009) Association study of IFNAR2 and IL10RB genes with the susceptibility and interferon response in HBV infection. *J Viral Hepat*, 16, 674-80.
- GUERIN, G., BAILEY, E., BERNOCO, D., ANDERSON, I., ANTCZAK, D. F., BELL, K., BIROS, I., BJORNSTAD, G., BOWLING, A. T., BRANDON, R., CAETANO, A. R., CHOLEWINSKI, G., COLLING, D., EGGLESTON, M., ELLIS, N., FLYNN, J., GRALAK, B., HASEGAWA, T., KETCHUM, M., LINDGREN, G., LYONS, L. A., MILLON, L. V., MARIAT, D., MURRAY, J., NEAU, A., ROED, K., SANDBERG, K., SKOW, L. C., TAMMEN, I., TOZAKI, T., VAN DYK, E., WEISS, B., YOUNG, A. & ZIEGLE, J. (2003) The second generation of the International Equine Gene Mapping Workshop half-sibling linkage map. *Anim Genet*, 34, 161-8.
- GUPTA, N., MARTIN, P. M., PRASAD, P. D. & GANAPATHY, V. (2006) SLC5A8 (SMCT1)-mediated transport of butyrate forms the basis for the tumor suppressive function of the transporter. *Life Sci*, 78, 2419-25.
- HARBERS, M. & CARNINCI, P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat Methods*, 2, 495-502.
- HARDY, G. H. (1908) Mendelian Proportions in a Mixed Population. *Science*, 28, 49-50.
- HAUSSINGER, D. (1996) The role of cellular hydration in the regulation of cell function. *Biochem J*, 313 (Pt 3), 697-710.
- HERT, D. G., FREDLAKE, C. P. & BARRON, A. E. (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 29, 4618-26.
-

-
- HINIKER, A., VERTOMMEN, D., BARDWELL, J. C. & COLLET, J. F. (2006) Evidence for conformational changes within DsbD: possible role for membrane-embedded proline residues. *J Bacteriol*, 188, 7317-20.
- HIRSCHHORN, J. N. & DALY, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6, 95-108.
- HODGES, E., XUAN, Z., BALIJA, V., KRAMER, M., MOLLA, M. N., SMITH, S. W., MIDDLE, C. M., RODESCH, M. J., ALBERT, T. J., HANNON, G. J. & MCCOMBIE, W. R. (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, 39, 1522-7.
- HOFFBRAND, A. V., PETTIT, J.E. AND MOSS, P.A.H. (2001) *Haematology*, Blackwell Publishing.
- JAIN, N. C. (1993) *Essentials of Veterinary Hematology.*, U.S.A., Williams and Wilkins.
- JELINEK, F., FALDYNA, M. & JASURKOVA-MIKUTOVA, G. (2006) Severe combined immunodeficiency in a Fell pony foal. *J Vet Med A Physiol Pathol Clin Med*, 53, 69-73.
- JIN, L., ZHU, W. & GUO, J. (2010) Genome-wide association studies using haplotype clustering with a new haplotype similarity. *Genet Epidemiol*, 34, 633-41.
- KAHRIZI, K., HU, C. H., GARSHASBI, M., ABEDINI, S. S., GHADAMI, S., KARIMINEJAD, R., ULLMANN, R., CHEN, W., ROPERS, H. H., KUSS, A. W., NAJMABADI, H. & TZSCHACH, A. (2011) Next generation sequencing in a family with autosomal recessive Kahrizi syndrome (OMIM 612713) reveals a homozygous frameshift mutation in SRD5A3. *Eur J Hum Genet*, 19, 115-7.
- KARLSSON, E. K., BARANOWSKA, I., WADE, C. M., SALMON HILLBERTZ, N. H., ZODY, M. C., ANDERSON, N., BIAGI, T. M., PATTERSON, N., PIELBERG, G. R., KULBOKAS, E. J., 3RD, COMSTOCK, K. E., KELLER, E. T., MESIROV, J. P., VON EULER, H., KAMPE, O., HEDHAMMAR, A., LANDER, E. S., ANDERSSON, G., ANDERSSON, L. & LINDBLAD-TOH, K. (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*, 39, 1321-8.
- KINO, T., SEGARS, J. H. & CHROUSOS, G. P. (2010) The Guanine Nucleotide Exchange Factor Brx: A Link between Osmotic Stress, Inflammation and
-

-
- Organ Physiology and Pathophysiology. *Expert Rev Endocrinol Metab*, 5, 603-614.
- KREMYANSKAYA, M. & MONROE, J. G. (2005) Ig-independent Ig beta expression on the surface of B lymphocytes after B cell receptor aggregation. *J Immunol*, 174, 1501-6.
- KWON, H. M., YAMAUCHI, A., UCHIDA, S., PRESTON, A. S., GARCIA-PEREZ, A., BURG, M. B. & HANDLER, J. S. (1992) Cloning of the cDNA for a Na⁺/myo-inositol cotransporter, a hypertonicity stress protein. *J Biol Chem*, 267, 6297-301.
- LAAN, T. T., GOEHRING, L. S. & SLOET VAN OLDRUITENBORGH-OOSTERBAAN, M. M. (2005) Von Willebrand's disease in an eight-day-old quarter horse foal. *Vet Rec*, 157, 322-4.
- LAIRD, N. M. & LANGE, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*, 7, 385-94.
- LAKATOS, P. L., KISS, L. S., PALATKA, K., ALTORJAY, I., ANTAL-SZALMAS, P., PALYU, E., UDVARDY, M., MOLNAR, T., FARKAS, K., VERES, G., HARSFALVI, J., PAPP, J. & PAPP, M. (2011) Serum lipopolysaccharide-binding protein and soluble CD14 are markers of disease activity in patients with Crohn's disease. *Inflamm Bowel Dis*, 17, 767-77.
- LANDER, E. S. & BOTSTEIN, D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, 236, 1567-70.
- LEWIS, C. M. (2002) Genetic association studies: design, analysis and interpretation. *Brief Bioinform*, 3, 146-53.
- LIU, N., ZHANG, K. & ZHAO, H. (2008) Haplotype-association analysis. *Adv Genet*, 60, 335-405.
- MAKITIE, O., SULISALO, T., DE LA CHAPELLE, A. & KAITILA, I. (1995) Cartilage-hair hypoplasia. *J Med Genet*, 32, 39-43.
- MALLEE, J. J., ATTA, M. G., LORICA, V., RIM, J. S., KWON, H. M., LUCENTE, A. D., WANG, Y. & BERRY, G. T. (1997) The structural organization of the human Na⁺/myo-inositol cotransporter (SLC5A3) gene and characterization of the promoter. *Genomics*, 46, 459-65.
-

-
- MARDIS, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24, 133-41.
- MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. & GILAD, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18, 1509-17.
- MCCUE, M. E., VALBERG, S. J., MILLER, M. B., WADE, C., DIMAURO, S., AKMAN, H. O. & MICKELSON, J. R. (2008) Glycogen synthase (GYS1) mutation causes a novel skeletal muscle glycogenosis. *Genomics*, 91, 458-66.
- MCGUIRE, T. C. & POPPIE, M. J. (1973) Hypogammaglobulinemia and thymic hypoplasia in horses: a primary combined immunodeficiency disorder. *Infect Immun*, 8, 272-7.
- MELAMEDE, M. (1985) Automatable process for sequencing nucleotide. . U.S.A.
- MICKELSON, J. (2011) Update on new genome tools: Design of Equine 74K SNP chip. *International Plant and Animal Genome XIX*. San Diego, California.
- MITTMANN, E. H., LAMPE, V., MOMKE, S., ZEITZ, A. & DISTL, O. (2010) Characterization of a minimal microsatellite set for whole genome scans informative in warmblood and coldblood horse breeds. *J Hered*, 101, 246-50.
- MOROZOVA, O. & MARRA, M. A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92, 255-64.
- MORRISSY, S., ZHAO, Y., DELANEY, A., ASANO, J., DHALLA, N., LI, I., MCDONALD, H., PANDOH, P., PRABHU, A. L., TAM, A., HIRST, M. & MARRA, M. (2010) Digital gene expression by tag sequencing on the illumina genome analyzer. *Curr Protoc Hum Genet*, Chapter 11, Unit 11 11 1-36.
- MORTON, N. E. (1955) Sequential tests for the detection of linkage. *Am J Hum Genet*, 7, 277-318.
- MOSKVIN, V. & SCHMIDT, K. M. (2008) On multiple-testing correction in genome-wide association studies. *Genet Epidemiol*, 32, 567-73.
- NAGALAKSHMI, U., WAERN, K. & SNYDER, M. (2010) RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol*, Chapter 4, Unit 4 11 1-13.
-

-
- NIKOPOULOS, K., GILISSEN, C., HOISCHEN, A., VAN NOUHUYS, C. E., BOONSTRA, F. N., BLOKLAND, E. A., ARTS, P., WIESKAMP, N., STROM, T. M., AYUSO, C., TILANUS, M. A., BOUWHUIS, S., MUKHOPADHYAY, A., SCHEFFER, H., HOEFSLOOT, L. H., VELTMAN, J. A., CREMERS, F. P. & COLLIN, R. W. (2010) Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet*, 86, 240-7.
- PAHL, R. & SCHAFFER, H. (2010) PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics*, 26, 2093-100.
- PAN, Q., SHAI, O., LEE, L. J., FREY, B. J. & BLENCOWE, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40, 1413-5.
- PARKER, H. G., KUKEKOVA, A. V., AKEY, D. T., GOLDSTEIN, O., KIRKNESS, E. F., BAYSAC, K. C., MOSHER, D. S., AGUIRRE, G. D., ACLAND, G. M. & OSTRANDER, E. A. (2007) Breed relationships facilitate fine-mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. *Genome Res*, 17, 1562-71.
- PATTERSON, E. E., MINOR, K. M., TCHERNATYNSKAIA, A. V., TAYLOR, S. M., SHELTON, G. D., EKENSTEDT, K. J. & MICKELSON, J. R. (2008) A canine DNMI mutation is highly associated with the syndrome of exercise-induced collapse. *Nat Genet*, 40, 1235-9.
- PERRYMAN, L. E. (2000) Primary immunodeficiencies of horses. *Vet Clin North Am Equine Pract*, 16, 105-16, vii.
- PERRYMAN, L. E., MCGUIRE, T. C. & BANKS, K. L. (1983) Animal model of human disease. Infantile X-linked agammaglobulinemia. Agammaglobulinemia in horses. *Am J Pathol*, 111, 125-7.
- PERRYMAN, L. E., MCGUIRE, T. C. & HILBERT, B. J. (1977) Selective immunoglobulin M deficiency in foals. *J Am Vet Med Assoc*, 170, 212-5.
- PERRYMAN, L. E., MCGUIRE, T. C. & TORBECK, R. L. (1980) Ontogeny of lymphocyte function in the equine fetus. *Am J Vet Res*, 41, 1197-200.
- PETERFY, M., MAO, H. Z. & DOOLITTLE, M. H. (2006) The cld mutation: narrowing the critical chromosomal region and selecting candidate genes. *Mamm Genome*, 17, 1013-24.
-

-
- POHLENZ, J. & REFETOFF, S. (1999) Mutations in the sodium/iodide symporter (NIS) gene as a cause for iodide transport defects and congenital hypothyroidism. *Biochimie*, 81, 469-76.
- POLMAR, S. H. & PIERCE, G. F. (1986) Cartilage hair hypoplasia: immunological aspects and their clinical implications. *Clin Immunol Immunopathol*, 40, 87-93.
- PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38, 904-9.
- PRICE, A. L., ZAITLEN, N. A., REICH, D. & PATTERSON, N. (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11, 459-463.
- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81, 559-75.
- RATHGEBER, R. A., BROOKS, M. B., BAIN, F. T. & BYARS, T. D. (2001) Clinical vignette. Von Willebrand disease in a Thoroughbred mare and foal. *J Vet Intern Med*, 15, 63-6.
- REHMAN, A. U., MORELL, R. J., BELYANTSEVA, I. A., KHAN, S. Y., BOGER, E. T., SHAHZAD, M., AHMED, Z. M., RIAZUDDIN, S., KHAN, S. N. & FRIEDMAN, T. B. (2010) Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet*, 86, 378-88.
- RICHARDS, A. J., KELLY, D. F., KNOTTENBELT, D. C., CHEESEMAN, M. T. & DIXON, J. B. (2000) Anaemia, diarrhoea and opportunistic infections in Fell ponies. *Equine Vet J*, 32, 386-91.
- RONAGHI, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res*, 11, 3-11.
- ROSENKRANZ, R., BORODINA, T., LEHRACH, H. & HIMMELBAUER, H. (2008) Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics*, 92, 187-94.
-

-
- SANGER, F., NICKLEN, S. & COULSON, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74, 5463-7.
- SANTER, R., RISCHEWSKI, J., BLOCK, G., KINNER, M., WENDEL, U., SCHAUB, J. & SCHNEPPENHEIM, R. (2000) Molecular analysis in glycogen storage disease 1 non-A: DHPLC detection of the highly prevalent exon 8 mutations of the G6PT1 gene in German patients. *Hum Mutat*, 16, 177.
- SCHOLES, S. F., HOLLIMAN, A., MAY, P. D. & HOLMES, M. A. (1998) A syndrome of anaemia, immunodeficiency and peripheral ganglionopathy in Fell pony foals. *Vet Rec*, 142, 128-34.
- SCHROEDER, A., MUELLER, O., STOCKER, S., SALOWSKY, R., LEIBER, M., GASSMANN, M., LIGHTFOOT, S., MENZEL, W., GRANZOW, M. & RAGG, T. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol*, 7, 3.
- SCHUELKE, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol*, 18, 233-4.
- SEELow, D., SCHWARZ, J. M. & SCHUELKE, M. (2008) GeneDistiller--distilling candidate genes from linkage intervals. *PLoS One*, 3, e3874.
- SELLON, D. C. (2000) Secondary immunodeficiencies of horses. *Vet Clin North Am Equine Pract*, 16, 117-30.
- SHIRAKI, T., KONDO, S., KATAYAMA, S., WAKI, K., KASUKAWA, T., KAWAJI, H., KODZIUS, R., WATAHIKI, A., NAKAMURA, M., ARAKAWA, T., FUKUDA, S., SASAKI, D., PODHAJSKA, A., HARBERS, M., KAWAI, J., CARNINCI, P. & HAYASHIZAKI, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100, 15776-81.
- SILBERSTEIN, M., TZEMACH, A., DOVGOLEVSKY, N., FISHELSON, M., SCHUSTER, A. & GEIGER, D. (2006) Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *Am J Hum Genet*, 78, 922-35.
- SINGER, V. L., JONES, L. J., YUE, S. T. & HAUGLAND, R. P. (1997) Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. *Anal Biochem*, 249, 228-38.
-

-
- SPENCER, C. C., SU, Z., DONNELLY, P. & MARCHINI, J. (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5, e1000477.
- SWINBURNE, J., GERSTENBERG, C., BREEN, M., ALDRIDGE, V., LOCKHART, L., MARTI, E., ANTCZAK, D., EGGLESTON-STOTT, M., BAILEY, E., MICKELSON, J., ROED, K., LINDGREN, G., VON HAERINGEN, W., GUERIN, G., BJARNASON, J., ALLEN, T. & BINNS, M. (2000) First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families. *Genomics*, 66, 123-34.
- SWINBURNE, J., LOCKHART, L., SCOTT, M. & BINNS, M. M. (1999) Estimation of the prevalence of severe combined immunodeficiency disease in UK Arab horses as determined by a DNA-based test. *Vet Rec*, 145, 22-3.
- SWINBURNE, J. E., BOGLE, H., KLUKOWSKA-ROTZLER, J., DROGEMULLER, M., LEEB, T., TEMPERTON, E., DOLF, G. & GERBER, V. (2009) A whole-genome scan for recurrent airway obstruction in Warmblood sport horses indicates two positional candidate regions. *Mamm Genome*, 20, 504-15.
- SWINBURNE, J. E., BOURSNELL, M., HILL, G., PETTITT, L., ALLEN, T., CHOWDHARY, B., HASEGAWA, T., KUROSAWA, M., LEEB, T., MASHIMA, S., MICKELSON, J. R., RAUDSEPP, T., TOZAKI, T. & BINNS, M. (2006) Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families. *Genomics*, 87, 1-29.
- SWINBURNE, J. E., HOPKINS, A. & BINNS, M. M. (2002) Assignment of the horse grey coat colour gene to ECA25 using whole genome scanning. *Anim Genet*, 33, 338-42.
- TERRY, R. B., ARCHER, S., BROOKS, S., BERNOCO, D. & BAILEY, E. (2004) Assignment of the appaloosa coat colour gene (LP) to equine chromosome 1. *Anim Genet*, 35, 134-7.
- THOMAS, G. W. (2003) Immunodeficiency in Fell Ponies. *Veterinary Pathology*. Liverpool, University of Liverpool.
- THOMAS, G. W., BELL, S. C. & CARTER, S. D. (2005) Immunoglobulin and peripheral B-lymphocyte concentrations in Fell pony foal syndrome. *Equine Vet J*, 37, 48-52.
-

-
- TOZAKI, T., SWINBURNE, J., HIROTA, K., HASEGAWA, T., ISHIDA, N. & TOBE, T. (2007) Improved resolution of the comparative horse-human map: investigating markers with in silico and linkage mapping approaches. *Gene*, 392, 181-6.
- TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25, 1105-11.
- TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28, 511-5.
- TRINH-TRANG-TAN, M. M., VILELA-LAMEGO, C., PICOT, J., WAUTIER, M. P. & CARTRON, J. P. (2010) Intercellular adhesion molecule-4 and CD36 are implicated in the abnormal adhesiveness of sickle cell SAD mouse erythrocytes to endothelium. *Haematologica*, 95, 730-7.
- TRYON, R. C., WHITE, S. D. & BANNASCH, D. L. (2007) Homozygosity mapping approach identifies a missense mutation in equine cyclophilin B (PIB) associated with HERDA in the American Quarter Horse. *Genomics*, 90, 93-102.
- TURK, E., ZABEL, B., MUNDLOS, S., DYER, J. & WRIGHT, E. M. (1991) Glucose/galactose malabsorption caused by a defect in the Na⁺/glucose cotransporter. *Nature*, 350, 354-6.
- TURRENTINE, M. A., SCULLEY, P. W., GREEN, E. M. & JOHNSON, G. S. (1986) Prekallikrein deficiency in a family of miniature horses. *Am J Vet Res*, 47, 2464-7.
- TZENG, J. Y., WANG, C. H., KAO, J. T. & HSIAO, C. K. (2006) Regression-based association analysis with clustered haplotypes through use of genotypes. *Am J Hum Genet*, 78, 231-42.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. & KINZLER, K. W. (1995) Serial analysis of gene expression. *Science*, 270, 484-7.
- VERMEER, S., HOISCHEN, A., MEIJER, R. P., GILISSEN, C., NEVELING, K., WIESKAMP, N., DE BROUWER, A., KOENIG, M., ANHEIM, M., ASSOUM, M., DROUOT, N., TODOROVIC, S., MILIC-RASIC, V., LOCHMULLER, H., STEVANIN, G., GOIZET, C., DAVID, A., DURR, A., BRICE, A., KREMER, B., VAN DE WARRENBURG, B. P.,
-

-
- SCHIJVENAARS, M. M., HEISTER, A., KWINT, M., ARTS, P., VAN DER WIJST, J., VELTMAN, J., KAMSTEEG, E. J., SCHEFFER, H. & KNOERS, N. (2010) Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals ANO10 mutations in patients with autosomal-recessive cerebellar ataxia. *Am J Hum Genet*, 87, 813-9.
- VILSEN, B., ANDERSEN, J. P., CLARKE, D. M. & MACLENNAN, D. H. (1989) Functional consequences of proline mutations in the cytoplasmic and transmembrane sectors of the Ca²⁺(+)-ATPase of sarcoplasmic reticulum. *J Biol Chem*, 264, 21024-30.
- WADE, C. M., GIULOTTO, E., SIGURDSSON, S., ZOLI, M., GNERRE, S., IMSLAND, F., LEAR, T. L., ADELSON, D. L., BAILEY, E., BELLONE, R. R., BLOCKER, H., DISTL, O., EDGAR, R. C., GARBER, M., LEEB, T., MAUCELI, E., MACLEOD, J. N., PENEDO, M. C., RAISON, J. M., SHARPE, T., VOGEL, J., ANDERSSON, L., ANTCZAK, D. F., BIAGI, T., BINNS, M. M., CHOWDHARY, B. P., COLEMAN, S. J., DELLA VALLE, G., FRYC, S., GUERIN, G., HASEGAWA, T., HILL, E. W., JURKA, J., KHALAINEN, A., LINDGREN, G., LIU, J., MAGNANI, E., MICKELSON, J. R., MURRAY, J., NERGADZE, S. G., ONOFRIO, R., PEDRONI, S., PIRAS, M. F., RAUDSEPP, T., ROCCHI, M., ROED, K. H., RYDER, O. A., SEARLE, S., SKOW, L., SWINBURNE, J. E., SYVANEN, A. C., TOZAKI, T., VALBERG, S. J., VAUDIN, M., WHITE, J. R., ZODY, M. C., LANDER, E. S. & LINDBLAD-TOH, K. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326, 865-7.
- WANG, Z., GERSTEIN, M. & SNYDER, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10, 57-63.
- WARD, A. C., TOUW, I. & YOSHIMURA, A. (2000) The Jak-Stat pathway in normal and perturbed hematopoiesis. *Blood*, 95, 19-29.
- WELDON, A. D., ZHANG, C., ANTCZAK, D. F. & REBHUN, W. C. (1992) Selective IgM deficiency and abnormal B-cell response in a foal. *J Am Vet Med Assoc*, 201, 1396-8.
- WILHELM, B. T. & LANDRY, J. R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48, 249-57.
- WILHELM, B. T., MARGUERAT, S., GOODHEAD, I. & BAHLER, J. (2010) Defining transcribed regions using RNA-seq. *Nat Protoc*, 5, 255-66.
-

- WRIGHT, E. M. & TURK, E. (2004) The sodium/glucose cotransport family SLC5. *Pflugers Arch*, 447, 510-8.
- YUE, W., YANG, Y., ZHANG, Y., LU, T., HU, X., WANG, L., RUAN, Y., LV, L. & ZHANG, D. (2011) A case-control association study of NRXN1 polymorphisms with schizophrenia in Chinese Han population. *Behav Brain Funct*, 7, 7.

Supporting papers

1. Fox-Clipsham, L., Swinburne, J.E., Papoula-Pereira, R.I., Blunden, A.S., Malalana, F., Knottenbelt, D.C., Carter, S.D (2009) Immunodeficiency/ anaemia syndrome in a Dales pony. *Vet Rec*, 165(10):289-90.
 2. Fox-Clipsham L.Y., Carter, S.D., Goodhead, I., Hall, N., Knottenbelt, D.C., May, P.D.F., Ollier, W.E. and Swinburne, J.E. (2011) Identification of a Mutation in SLC5A3 Related to Fatal Foal Immunodeficiency Syndrome in the Fell and Dales Pony. *PLoS Genet*, 7(7): (Epub 2011 Jul 7).
 3. Fox-Clipsham L.Y., Brown E.E., Carter S.D. and Swinburne J.E. (2011) Population screening of endangered horse breeds for the foal immunodeficiency syndrome mutation. *Vet Rec*, Oct 20: (Epub ahead of print).
-
-

SHORT COMMUNICATIONS

Immunodeficiency/anaemia syndrome in a Dales pony

L. Fox-Clipsham, J. E. Swinburne, R. I. Papoula-Pereira,
A. S. Blunden, F. Malalana, D. C. Knottenbelt, S. D. Carter

THE syndrome of anaemia and immunodeficiency was first recognised in Fell pony foals in the UK in 1997 (Scholes and others 1998) and has since been reported in the same breed in the Netherlands (Butler and others 2006) and the USA (Gardner and others 2006); there have been no reports of the syndrome in any other horse breed. Affected foals are apparently normal at birth, but the disease first manifests at two to six weeks of age; the reported characteristic clinical signs include weakness, dyspnoea, nasal discharge, poor growth, reduced appetite, diarrhoea and pale gums. A profound and progressive fall in red blood cell count (packed cell value [PCV] <20 per cent) is a notable early feature. The number of circulating lymphocytes is reduced and there is an increase in the number of peripheral blood polymorphonuclear cells (Dixon and others 2000). Analyses of lymphocyte subpopulations show normal numbers of circulating T lymphocytes (Bell and others 2001) but severely reduced numbers of circulating B lymphocytes (Thomas and others 2003); there are also low concentrations of circulating immunoglobulins in affected foals (Thomas and others 2005). These changes coincide with episodes of opportunistic bacterial, viral and parasitic infections. Typically, these changes persist for three to six weeks, with the foal becoming progressively weaker due to systemic infections and profound anaemia; the PCV can decrease to as low as 3 per cent.

Attempts at symptomatic treatment have been made; including rehydration, antibiotics, enteral/parenteral nutrition, analgesia, supplementation with vitamins and selenium, blood transfusions and injections of erythropoietin. None of these is reported to have had any effect apart from a short-term delay in the death or euthanasia of the affected foal. No foals have been recorded as surviving the syndrome.

At postmortem examination, there are many distinctive changes associated with the disorder, including glossal hyperkeratosis and diphtheritic fungal glossitis (usually associated with *Candida* species), colitis, pneumonia, inactive or aplastic bone marrow, and a hypoplastic thymus and lymph nodes (Scholes and others 1998, Richards and others 2000).

Studbook analysis has clearly indicated that the disease has a very strong genetic component, and the heritability pattern is consistent with it being an autosomal recessive condition. The unusual husbandry of Fell ponies, in which herds of ponies may be kept free-living on the upland areas of the northern UK moors, means it is difficult to make estimates of the annual incidence of the disease. However, it probably affects at least 5 to 10 per cent of foals born each year in the past 10 years, which indicates a high penetration rate of a mutant gene in the breeding population. Further analysis of the studbook indicates that carriers of the disease in the breed can be traced back to the early 1950s, and also, during the history of the breed, there have been frequent crossbreedings recorded with other breeds, most notably the Dales pony. Until 2008, there had been no reports of Fell pony syndrome in any other breed of horse, but the silent nature of the gene penetration seen in autosomal recessive conditions may have allowed the mutation to be inherited and spread in another breed before any cases were reported.

This short communication reports the first definitive diagnosis of a case of anaemia/immunodeficiency syndrome in another breed of horse, the Dales pony.

A female Dales pony foal was born in May 2008 in the UK; the sire and dam were both registered Dales ponies. The foal was apparently normal at birth and suckled successfully, but at 15 days of age it was presented to the local veterinary surgeon with diarrhoea and nasal discharge. A sample of the foal's plasma was tested for maternally derived antibodies, with positive results (IgG >8 g/l). The foal was treated with oral ceftiofur (Excenel; Pfizer) for five days, with no apparent improvement. Another five-day course of ceftiofur was administered. By 25 days of age, the nasal discharge had become more severe; an intestinal protectant (Diarsanyl; CEVA Animal Health) was administered in an attempt to control the diarrhoea, and an oral electrolyte solution (Lectade; Pfizer) was administered to treat dehydration. The foal was fed 400 ml of mare's milk over two hours twice via a stomach tube, alternating with 400 ml of the electrolyte solution.

At this point the foal was coughing and having trouble swallowing any liquids. Until 26 days of age, its body temperature had remained between 38.3 and 38.5°C; on day 26 this rose to 39.7°C and later on the same day to 39.9°C. The foal was referred to the Large Animal Hospital at the Faculty of Veterinary Science, University of Liverpool, for investigation of a possible immunocompromising disorder.

On admission, at 31 days of age, the foal was bright but pyrexial (39.3°C) and had obvious adventitious crackles on thoracic auscultation. Biochemistry was unremarkable apart from hypokalaemia (1.5 mmol/l). The foal was given intravenous fluids supplemented with 20 mmol/l potassium chloride. Haematology revealed marked anaemia (PCV 14 per cent) with leucocytosis (white blood cell count 15.6×10^9 cells/l). On day 34, a haematology profile showed a white blood cell count of 5.33×10^9 cells/l, consisting of 43 per cent lymphocytes, 47 per cent neutrophils and 10 per cent monocytes. A blood smear showed the anaemia to be non-regenerative. A Coombs test for anti-red cell antibodies was negative. Examination of a faecal sample showed the presence of *Cryptosporidium* species oocysts.

Thoracic radiographs showed consolidation of the ventral lung lobes, and culture of a transcutaneous tracheal wash revealed *Candida albicans*. Endoscopy of the airways revealed no abnormalities. The foal showed a minor improvement following treatment with 1 mg/kg cefquinome (Cephaguard; Intervet), administered intravenously twice a day, 20 mg/kg benzylpenicillin (Crystapen; Intervet/Schering-Plough Animal Health), administered intravenously four times a day, and 4 mg omeprazole (GastroGard; Merial) administered orally once a day, including improved suckling from the mare, but the anaemia progressed and the PCV fell to 10.0 per cent on day 39 and subsequently to 7.9 per cent on day 41. Following discussion with the owner, the foal was euthanased and tissue samples were taken immediately for histological analyses.

Veterinary Record (2009) **165**, 289-290

L. Fox-Clipsham, BSc,
J. E. Swinburne, BSc, PhD,
A. S. Blunden, BVetMed, PhD,
FRCPath, MRCVS,
Animal Health Trust, Lanwades Park,
Kentford, Newmarket, Suffolk
CB8 7UU

R. I. Papoula-Pereira, LMV,
F. Malalana, DVM, MRCVS,
D. C. Knottenbelt, BVM&S, DVMS,
DipECEIM, MRCVS,
Department of Veterinary Clinical
Science, University of Liverpool,
Leahurst, Neston, Cheshire
CH64 7TE

S. D. Carter, BSc, PhD, FRCPath,
Department of Veterinary Pathology,
University of Liverpool, Brownlow
Hill, Liverpool L69 7ZJ

Correspondence to Professor Carter,
e-mail: scarter@liv.ac.uk

SHORT COMMUNICATIONS

On gross examination, the cranioventral lobes of the lungs had severe, focally extensive atelectasis; the remaining lobes had severe multifocal emphysema. In the small intestine, the ileal and jejunal mucosae were roughened, with multifocal areas of reddening. The large intestinal mucosa also showed multifocal areas of reddening. The liver was diffusely congested. Each ovary had a cyst (0.7 cm in diameter) at the cranial pole, which was filled with pale fluid. The cut surface of the left ovary exhibited multifocal cystic cavities. Examination of the lymphoreticular system revealed that the spleen was diffusely reddened and most lymph nodes were mildly to moderately enlarged, but the prescapular lymph nodes were markedly reduced in size. The femoral bone was diffusely dark red. The thymus was not identifiable, and only adipose tissue was seen in its place. The other organs were apparently normal.

Histopathological examination of the paraffin wax-embedded sections of each tissue revealed the following results. The airways of the lung were filled with necrotic debris, squamous cells and occasional rod-shaped bacteria. Multifocal degeneration of the epithelial cells was seen, and scattered cells contained basophilic intranuclear inclusion bodies, consistent with adenovirus infection. In some sections, severe diffuse atelectasis and severe multifocal emphysema were seen. The tongue displayed varying degrees of multifocal mucosal ulceration, with a neutrophil infiltrate extending on to the adjacent muscularis. The surfaces of the ulcers contained large numbers of Gram-negative cocci and Gram-positive coccibacilli together with fungal hyphae and spores (periodic acid-Schiff positive), consistent with infection with *Candida* species. The small intestine had low numbers of cryptosporidia on the apical surfaces of enterocytes in the duodenum. Extensive areas of the duodenum showed loss of villi, fibrosis, and severe mononuclear and neutrophilic infiltration and crypt dilation, occasionally forming crypt abscesses.

Histology of the lymphoid organs showed consistent deficiencies. The splenic follicles were moderately cellular or inapparent, with a low number of lymphocytes and macrophages in the red pulp. Lymph nodes displayed poor architecture and no follicular development; some lymph nodes were moderately depleted of lymphocytes but had a substantial number of medullary macrophages. The lymph nodes displayed substantial lymphoid depletion. Sections of bone marrow showed a predominance of adipocytes, evidence of hypoplasia and an increased myeloid:erythroid ratio. Bone marrow smears revealed erythroid hypoplasia with a reduction in mature elements.

Immunohistology showed a low number of CD3+ lymphocytes in the bone marrow and a moderate number in the splenic red pulp. The bronchial lymph node had a large number of CD3+ cells in the paracortex. Staining for the B lymphocyte differentiation factor PAX-5 showed expression by occasional cells in the bone marrow. In the spleen, multiple aggregates of PAX-5-positive cells were seen in small follicles. The bronchial lymph node showed only a small number of PAX-5-positive cells, and lymphoid follicles were not seen. Staining for the B lymphocyte antigen receptor CD79A confirmed the staining patterns seen with PAX-5, namely, poor B cell immunostaining in lymph node follicles and reduced staining in the spleen. There was expression of CD79A in the bone marrow, but primarily by immature cells.

The clinical presentation, disease progression, and haematological, microbiological and histopathological findings all suggested that the foal had the same anaemia/immunodeficiency syndrome as that reported in

Fell ponies in the UK and elsewhere. The profound anaemia, allied to the clinical signs of opportunistic infection and the absence of B lymphocyte activity in the lymphoid organs, suggested that this disease is present in another horse breed. There is evidence to suggest that the responsible genetic lesion has been transmitted from Fell ponies to the Dales pony, although this has not been proved. Using studbook records, the disease can be traced back from known affected Fell pony foals to a common ancestor in the 1950s. It is likely that the Dales pony acquired the mutation by crossbreeding between the two breeds after that date. The possibility that the disease derives from a common genetic linkage between the two breeds going back many years is considered very unlikely. Discussions with the Dales Pony Society suggest that the present case is the first to occur in this breed, as the society had been aware of the potential risk of the immunodeficiency syndrome for many years, and has been vigilant for it. It is therefore highly likely that few Dales ponies are carriers of the mutation. However, until a genetic test is developed to identify carriers, it will be possible to identify carriers only by the production of affected foals. Three of the authors (LF-C, JES and SDC) will address this issue in studies to identify the genetic lesion in both Fell and Dales ponies.

It is important that breeders, owners and veterinary surgeons dealing with Dales ponies are alert to this disease, and they should refer any suspected case, including cases in other breeds where similar signs are presented, to the research team at the Animal Health Trust or the Faculty of Veterinary Science at Liverpool University. There are several other circumstances in which young foals might be anaemic, but any young foal with a PCV of less than 20 per cent and relevant clinical signs should be considered at risk of the anaemia/immunodeficiency syndrome.

Acknowledgements

The Horse Trust supports the authors' research into the Fell pony immunodeficiency syndrome.

References

- BELL, S. C., SAVIDGE, C., TAYLOR, P., KNOTTENBELT, D. C. & CARTER, S. D. (2001) An immunodeficiency in Fell ponies: a preliminary study into cellular responses. *Equine Veterinary Journal* **33**, 687-692
- BUTLER, C. M., WESTERMANN, C. M., KOEMAN, J. P. & SLOET VAN OLDRIJTBORGH-OOSTERBAAN, M. M. (2006) The fell pony immunodeficiency syndrome also occurs in the Netherlands: a review and six cases. *Tijdschrift voor Diergeneeskunde* **131**, 114-118
- DIXON, J. B., SAVAGE, M., WATTRET, A., TAYLOR, P., ROSS, G., CARTER, S. D. & OTHERS (2000) Discriminant and multiple regression analysis of anemia and opportunistic infection in Fell pony foals. *Veterinary Clinical Pathology* **29**, 84-86
- GARDNER, R. B., HART, K. A., STOKOL, T., DIVERS, T. J. & FLAMINIO, M. J. (2006) Fell pony syndrome in a pony in North America. *Journal of Veterinary Internal Medicine* **20**, 198-203
- RICHARDS, A. J., KELLY, D. F., KNOTTENBELT, D. C., CHEESEMAN, M. T. & DIXON, J. B. (2000) Anaemia, diarrhoea and opportunistic infections in fell ponies. *Equine Veterinary Journal* **32**, 386-391
- SCHOLES, S. F., HOLLIMAN, A., MAY, P. D. & HOLMES, M. A. (1998) A syndrome of anaemia, immunodeficiency and peripheral ganglionopathy in fell pony foals. *Veterinary Record* **142**, 128-134
- THOMAS, G. W., BELL, S. C. & CARTER, S. D. (2005) Immunoglobulin and peripheral B-lymphocyte concentrations in Fell pony foal syndrome. *Equine Veterinary Journal* **37**, 48-52
- THOMAS, G. W., BELL, S. C., PHYTHIAN, C., TAYLOR, P., KNOTTENBELT, D. C. & CARTER, S. D. (2003) Aid of the antemortem diagnosis of fell pony foal syndrome by the analysis of B lymphocytes. *Veterinary Record* **152**, 618-621

Identification of a Mutation Associated with Fatal Foal Immunodeficiency Syndrome in the Fell and Dales Pony

Laura Y. Fox-Clipsham¹, Stuart D. Carter², Ian Goodhead³, Neil Hall³, Derek C. Knottenbelt⁴, Paul D. F. May⁵, William E. Ollier⁶, June E. Swinburne^{1*}

1 Animal Health Trust, Newmarket, Suffolk, United Kingdom, **2** Department of Infection Biology, School of Veterinary Science, University of Liverpool, Liverpool, United Kingdom, **3** Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom, **4** Department of Veterinary Clinical Science, Equine Hospital, University of Liverpool, Liverpool, United Kingdom, **5** Townhead Veterinary Centre, Townhead Farm, Penrith, United Kingdom, **6** Centre for Integrated Genomic Medical Research, University of Manchester, Manchester, United Kingdom

Abstract

The Fell and Dales are rare native UK pony breeds at risk due to falling numbers, in-breeding, and inherited disease. Specifically, the lethal Mendelian recessive disease Foal Immunodeficiency Syndrome (FIS), which manifests as B-lymphocyte immunodeficiency and progressive anemia, is a substantial threat. A significant percentage (~10%) of the Fell ponies born each year dies from FIS, compromising the long-term survival of this breed. Moreover, the likely spread of FIS into other breeds is of major concern. Indeed, FIS was identified in the Dales pony, a related breed, during the course of this work. Using a stepwise approach comprising linkage and homozygosity mapping followed by haplotype analysis, we mapped the mutation using 14 FIS-affected, 17 obligate carriers, and 10 adults of unknown carrier status to a ~1 Mb region (29.8 – 30.8 Mb) on chromosome (ECA) 26. A subsequent genome-wide association study identified two SNPs on ECA26 that showed genome-wide significance after Bonferroni correction for multiple testing: BIEC2-692674 at 29.804 Mb and BIEC2-693138 at 32.19 Mb. The associated region spanned 2.6 Mb from ~29.6 Mb to 32.2 Mb on ECA26. Re-sequencing of this region identified a mutation in the sodium/myo-inositol cotransporter gene (*SLC5A3*); this causes a P446L substitution in the protein. This gene plays a crucial role in the regulatory response to osmotic stress that is essential in many tissues including lymphoid tissues and during early embryonic development. We propose that the amino acid substitution we identify here alters the function of *SLC5A3*, leading to erythropoiesis failure and compromise of the immune system. FIS is of significant biological interest as it is unique and is caused by a gene not previously associated with a mammalian disease. Having identified the associated gene, we are now able to eradicate FIS from equine populations by informed selective breeding.

Citation: Fox-Clipsham LY, Carter SD, Goodhead I, Hall N, Knottenbelt DC, et al. (2011) Identification of a Mutation Associated with Fatal Foal Immunodeficiency Syndrome in the Fell and Dales Pony. *PLoS Genet* 7(7): e1002133. doi:10.1371/journal.pgen.1002133

Editor: Samantha A. Brooks, Cornell University, United States of America

Received: February 9, 2011; **Accepted:** May 3, 2011; **Published:** July 7, 2011

Copyright: © 2011 Fox-Clipsham et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by The Horse Trust (Grant Number G808; <http://www.horsetrust.org.uk/>), who also supported LYF-C. Support was also provided by Fell Pony 2000, the Petplan Charitable Trust (<http://www.petplantrust.org/>), and the Dorothy Russell Havemeyer Foundation (<http://www.havemeyerfoundation.org/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: june.swinburne@ahtr.org.uk

Introduction

The Fell and Dales are related sturdy pony breeds traditionally used as pack animals to carry goods over the difficult upland terrain of northern England. Both breeds experienced near extinction during WWII, and the current populations are descended from very few animals. It is likely that this genetic bottleneck, together with the use of prominent sires, was responsible for the emergence of a fatal Mendelian recessive disease, FIS, which currently affects up to 10% of Fell and 1% of Dales foals (data from UK breed societies). Both of these breeds are registered with the Rare Breeds Survival Trust due to their falling numbers, and important position in the UK's agricultural heritage.

FIS was first described in 1998 as a unique syndrome in which affected Fell foals develop diarrhoea, cough and fail to suckle [1]. Despite an initial response to treatment, the infections persist and were shown to be due to a primary B-cell deficiency [2] associated with reduced antibody production, with tested immunoglobulin isotypes including IgM, IgG_A, IgG_B and IgG(T) being significantly reduced [3]. Paradoxically, circulating T-lymphocyte numbers are

normal [4]. The reduced antibody levels in affected foals are consistent with an inability to generate an adaptive immune response, resulting in immunodeficiency once colostrum-derived immunoglobulin levels decrease at 3–6 weeks of age. This loss of maternally derived antibodies correlates with typical onset of FIS signs at 4–6 weeks. Concurrently, affected foals develop a non-hemolytic, non-regenerative progressive profound anemia [5], in itself severe enough to cause death and the main marker for euthanasia decisions by vets.

As a result of FIS, foals die or are humanely destroyed between 1–3 months of age, the disease being 100% fatal. In 2009, this condition was reported in the Dales breed [6]; it is likely that the mutation has passed between the breeds given the similarity between them and the practice of occasional interbreeding. The clinical and pathological findings for FIS are compatible with a primary defect of genetic origin [1], and this is supported by extensive genealogical studies [6,7]. FIS has a pattern of inheritance typical of an autosomal recessive disease, and the likely founder animal, which features in both the Fell and Dales studbooks, has been traced by pedigree analysis.

Author Summary

Foal Immunodeficiency Syndrome (FIS) is a genetic disease that affects two related British pony breeds, namely the Fell and the Dales. Foals with FIS appear to be normal at birth but within a few weeks develop evidence of infection such as diarrhoea, pneumonia, etc. The infections are resistant to treatment, and the foals die or are euthanized before three months of age. The foals also suffer from a severe progressive anemia. Being a recessive condition, the disease is difficult to control without a diagnostic DNA test to identify symptom-free carrier parents. Within the last few years the horse genome has been sequenced, and this has allowed the development of tools to identify genetic mutations in the horse at high resolution. In this article we demonstrate the use of these new tools to identify the location of the FIS mutation. The presumptive causal lesion was then identified by sequencing this region. This has enabled us to develop a test that can be used to identify carrier ponies, allowing breeders to avoid FIS in their foal crop.

Primary immunodeficiencies, which include depleted levels of lymphocytes and/or immunoglobulins, have previously been reported in the horse. The recessive defect 'severe combined immunodeficiency' (SCID), which is found in the Arabian breed, comprises a fatal deficiency in T- and B-lymphocyte numbers and function. The underlying lesion was found to be a 5 base-pair deletion in the gene coding DNA-dependent kinase, catalytic subunit DNA-PK_{CS} [8], a protein involved in V(D)J recombination required for adaptive immunity [9]. Like FIS foals, SCID foals have a markedly reduced thymus and have reduced numbers of germinal centers in secondary lymphoid organs [10]; unlike SCID foals however, FIS foals have apparently normal numbers of circulating T-cells [4]. Primary agammaglobinemia is rare in horses and comprises of a complete absence of immunoglobulin and reduced peripheral B-lymphocyte levels, with normal T-lymphocyte activity [11]. In this respect there is a similarity to FIS, however primary agammaglobinemia is only observed in males and is X-linked. Furthermore, profound anemia in combination with B-lymphopenia has not previously been reported in the horse or any other species, and as such FIS appears to be a unique disease process.

Here we report the mapping and identification of the genetic lesion that causes FIS. An initial scan using microsatellite markers identified the chromosome region responsible. The opportune production of a SNP chip, which utilized the SNP variants generated during the sequencing of the equine genome (<http://www.broadinstitute.org/mammals/horse>) then allowed a confirmatory association scan. This was followed by re-sequencing of the implicated region in order to identify the causal mutation.

Results

Microsatellite Analysis

A genome-wide microsatellite scan was performed on 41 individuals taken from five pedigrees of Fell ponies in which FIS was segregating (Figure S1), using a panel of 228 markers (Table S1). The data were examined both for loss of heterozygosity and for linkage (Table S2). Only one microsatellite, at 30.25 Mb on ECA26, showed a significant loss of heterozygosity ($\chi^2 = 7.15$, $P = 0.028$) and significant linkage (LOD score = 3.29 at $\theta = 0$) to the disease.

SNP Association Scan

The location of the lesion was confirmed and refined using a genome-wide association analysis with an equine SNP array (Illumina EquineSNP50 Infinium BeadChip), which contains

54,602 validated SNPs. After applying quality control (see Materials and Methods), data were available for 42,536 SNPs in 49 individuals (18 FIS-affected and 31 controls). To consider whether there was any population stratification among the samples, a multi-dimensional scaling plot of the genome-wide identity-by-state distances was performed (Figure S2); there was no significant difference between the affected and control samples for the first two components ($P = 0.553$). In addition, a quantile-quantile plot (Figure S3) to compare the expected and observed distributions of $-\log_{10}(P)$, obtained by a basic association test, showed that there was little evidence of inflation of the test statistics (genomic inflation factor $\lambda = 1.04$), indeed the test statistics appear to be marginally depressed rather than inflated. No correction was considered necessary. Two SNPs on ECA26 showed genome-wide significance after Bonferroni correction for multiple testing (Figure 1A). These are BIEC2-692674 at 29.804 Mb ($P_{\text{raw}} = 2.88 \times 10^{-7}$) and BIEC2-693138 at 32.19 Mb ($P_{\text{raw}} = 1.08 \times 10^{-6}$). The associated region spanned 2.6 Mb from position 29.6 Mb to 32.2 Mb on ECA26 (Figure 1B).

In a subsequent fine-mapping phase, 62 additional SNPs within the region were genotyped on 13 FIS-affected samples. Several novel SNPs were identified (dbSNP ss295469621-295469629). In addition, two further microsatellites were also genotyped (Table S1). The homozygous affected haplotype was shared by these animals over a 992 kb segment (Figure 2A). According to ENSEMBL gene prediction, fourteen genes lie in this interval (Figure 2B).

Resequencing of Critical Region

Five selected individual animals were re-sequenced over this critical region using sequence capture by NimbleGen arrays followed by GS FLX Titanium sequencing (GenBank submission under study accession no. ERP000492). The five individuals comprised one affected foal (A13 on Figure 2), the two obligate carrier parents, one apparently clear animal and one obligate carrier chosen for maximal homozygosity across the region. The FIS carrier status and familial relationships were confirmed for each individual by parentage verification. The last animal proved particularly useful in eliminating many potential causal variants. In total, eight verified SNPs were identified in the affected foal, narrowing the critical region to 375,063 bp (ECA26: 30,372,557 – 30,747,620 bp). Coverage of this critical interval was increased from 92.9% to 98.4% using Sanger sequencing; none of the remaining gaps fell within 200 bp of protein-coding sequence. Only one variant, a SNP at 30,660,224 bp, segregated as expected for a causal recessive mutation within the five sequenced samples. In addition there was no evidence of DNA rearrangement, duplication or insertion/deletion seen in the affected foal (Figure S4). The segregating SNP was assessed for validity as an FIS marker in equine populations.

Population Screen

Subsequently all 38 available affected foals (37 Fell, 1 Dales) were shown to be homozygous for the affected allele and all 21 available obligate carriers were heterozygous. A selection of Fell and Dales samples which were submitted to the Animal Health Trust for parentage verification between 2000 and 2010 were anonymously screened for the affected allele: 82 / 214 (38%) of the Fells and 16 / 87 (18%) of the Dales were heterozygous for the lesion and no homozygous affecteds were discovered. These carrier rates are consistent with the approximate observed disease prevalence of 10% in the Fell and 1% in the Dales populations. In addition, a selection of horse breeds (184 individuals from 11 breeds consisting of Thoroughbred ($n = 29$), Appaloosa ($n = 8$),

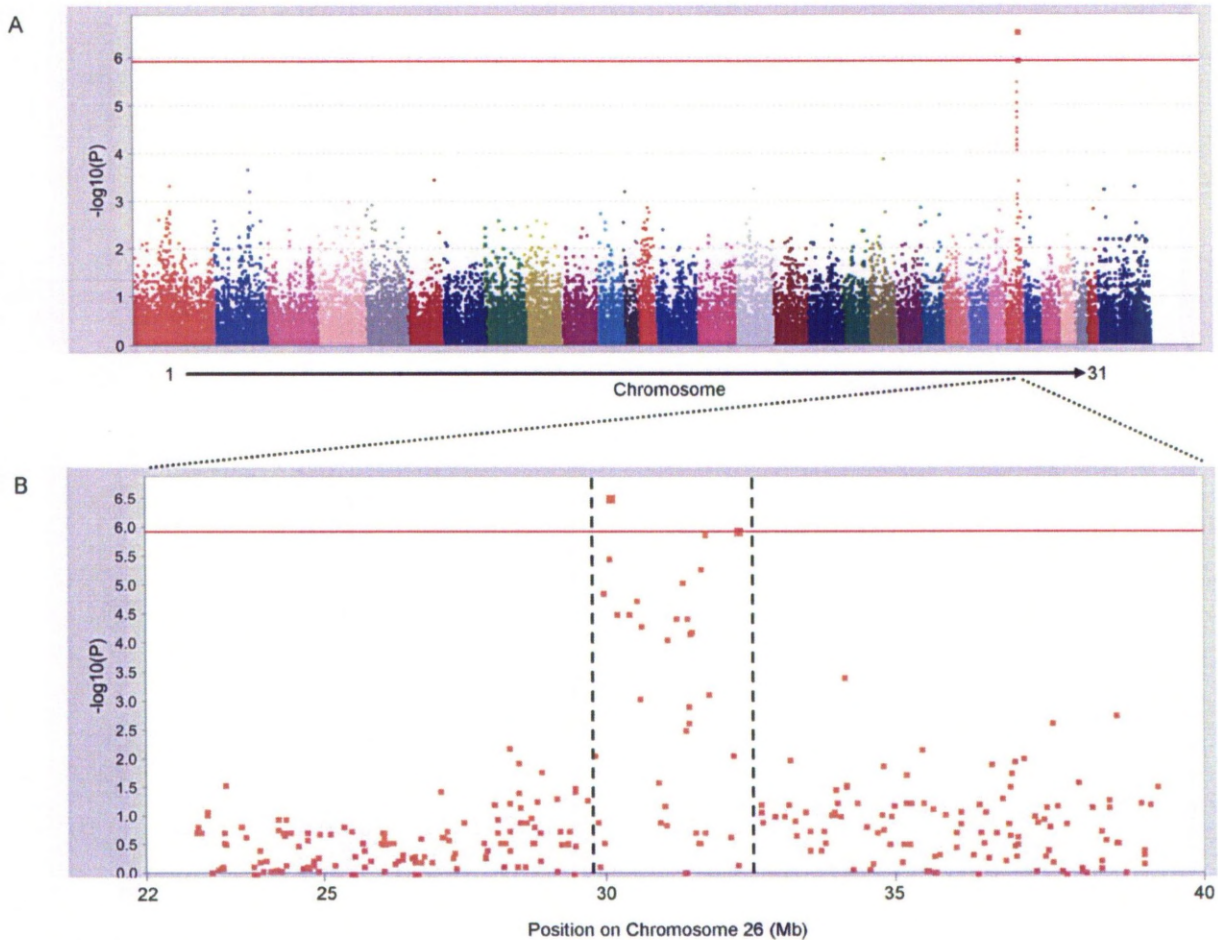


Figure 1. Identification of locus responsible for FIS by genome-wide association scan. (A) Manhattan plot of association test results showing the chromosome locations of the 42,536 SNPs which passed quality control against $-\log_{10}(P)$. The red line indicates the threshold for genome-wide significance after Bonferroni correction for multiple testing ($P_{\text{raw}} = 1.2 \times 10^{-6}$), which corresponds to an alpha value of 0.05. (B) Focus on region of ECA26 that shows FIS association. The vertical broken lines indicate the region (29.6–32.2 Mb) which was investigated further by fine-mapping.

doi:10.1371/journal.pgen.1002133.g001

Arab ($n = 21$), Warmblood Sport Horse ($n = 17$), Lipizzaner ($n = 2$), Cleveland Bay ($n = 20$), Dartmoor Pony ($n = 19$), Icelandic Horse ($n = 8$), New Forest Pony ($n = 20$), Shetland Pony ($n = 20$) and Shire ($n = 20$) which were considered unlikely to have interbred with either the Fell or Dales was genotyped and all proved homozygous wild-type.

Discussion

The identification of a mutation that segregates 100% with the disease has enabled a diagnostic test to be developed and offered to breeders and owners, allowing them to avoid carrier-carrier matings, and consequently drastically reduce the numbers of FIS-affected foals born each year. A gradual reduction in the use of carrier animals will, over time, lead to a reduction in the affected allele frequency in the population, while conserving the gene pool as much as possible. In addition, other equine breeds that may have interbred with the Fell or Dales will now be screened for FIS carriers.

The FIS-associated SNP falls within the single exon of the sodium/myo-inositol co-transporter gene (*SLC5A3*, also known as *SMIT*), which is a cell membrane transporter protein responsible for the co-transport of sodium ions and myo-inositol. This SNP is non-synonymous, causing a P446L substitution in *SLC5A3*; this amino acid residue (equivalent residue 451 in the human protein) is conserved in all 11 placental mammals for which high-coverage sequence is now available (selection shown in Figure 3). Similarly, this residue is conserved in other solute carrier family 5 (*SLC5*) paralogs in the horse which share similar structural homology (Figure 3).

The crystal structure of a bacterial homolog of *SLC5A1* (sodium/glucose co-transporter 1) has recently been elucidated [12] and shows this member of the *SLC5* family to have 14 transmembrane helices; the structural conformations adopted during transport, and the precise positions of substrate binding during transfer are now being identified [13]. Alignment of the protein sequences of the *SLC5* family suggests that P446 in equine *SLC5A3* is located in a transmembrane helix which is involved in

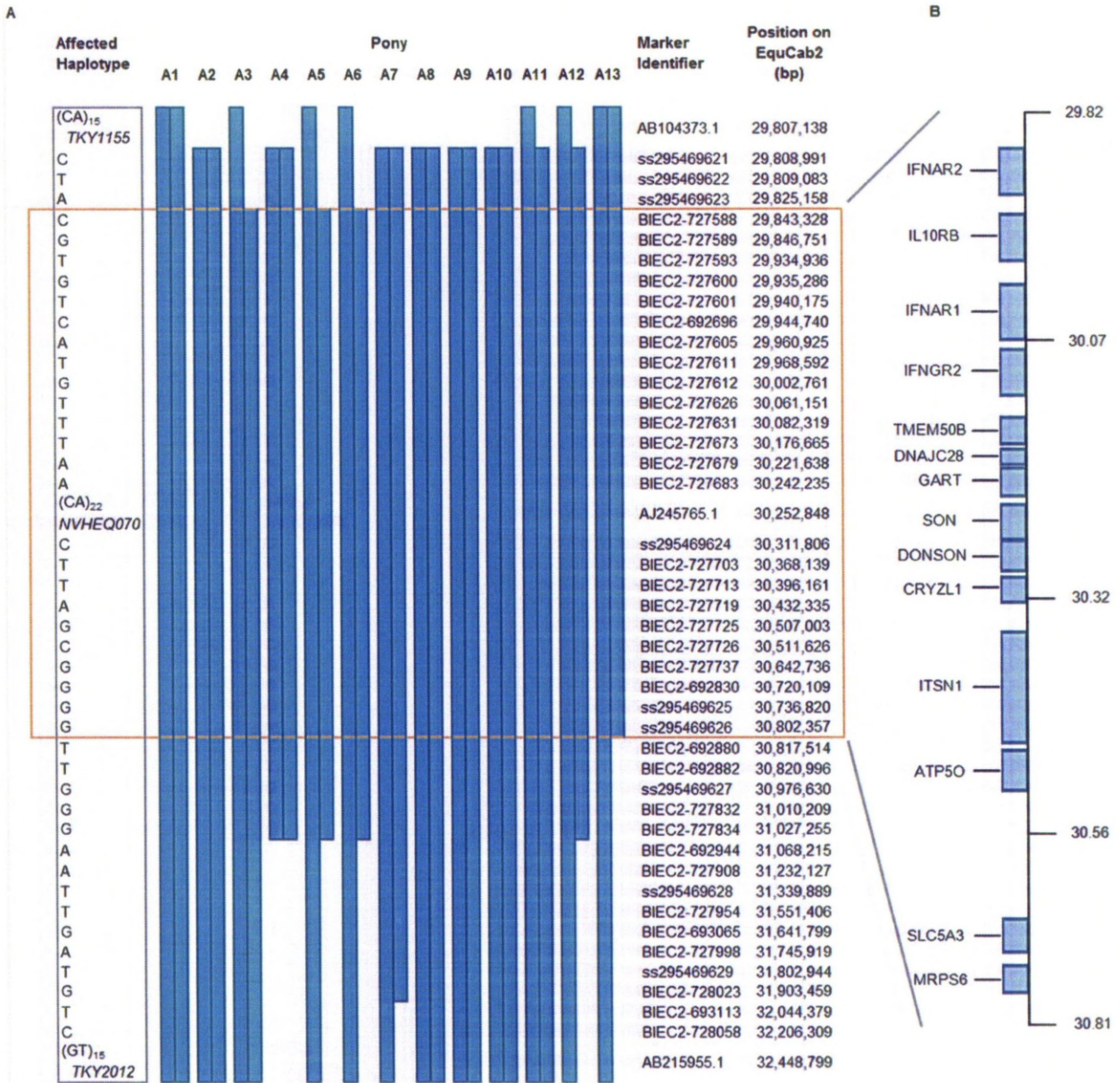


Figure 2. SNP haplotypes and genes from the region of ECA26 genetically linked and associated with FIS. (A) The affected SNP haplotype is shown on the left. The affected alleles for three key microsatellite markers are also included. The extent of conserved affected haplotypes present in the 13 affected individuals (A1–13) are indicated with blue bars. Accession numbers (newly identified SNPs) or the local SNP ID number (http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/), together with the genome position are given. The minimal 992 kb shared region of homozygosity (29,825 – 30,817 kb) is out-lined in orange. (B) The positions of the 14 ENSEMBL annotated genes within the conserved block are indicated.

doi:10.1371/journal.pgen.1002133.g002

forming the substrate cavity [12] and which tilts during substrate transfer. The two prolines at positions 445 and 446 may be required for effective substrate binding by closing the substrate binding site after the substrate is bound. Proline residues introduce structural destabilisation into alpha helices and have long been obvious candidates for points of conformational change required for substrate binding and release [14]. Indeed, replacement of prolines in the transmembrane helices of transport proteins has

shown that some residues are profoundly important in transport, affecting either substrate affinity or substrate movement [15].

SLC5A3 is an osmotic stress response gene, which acts to prevent dehydration caused by increased osmotic pressure in the extracellular environment. Dehydration causes the disruption of numerous cellular functions by denaturation of intracellular molecules and damage to sub-cellular architecture [16]. Extreme osmotic conditions are found in the kidney, although osmotic

Orthologs

Horse	SLC5A3 (FIS)	QMYLYIQEVADYLTPLVAALFLLAIFWKRCN
Horse	SLC5A3 (wt)	QMYLYIQEVADYLT PPVAALFLLAIFWKRCN
Human	SLC5A3	QMYLYIQEVADYLT PPVAALFLLAIFWKRCN
Mouse	SLC5A3	QMYLYIQEVADYLT PPVAALFLLAIFWKRCN
Dog	SLC5A3	QMYLYIQEVADYLT PPVAALFLLAIFWKRCN
Cow	SLC5A3	QMYLYIQEVADYLT PPVAALFLLAIFWKRCN

Paralogs

Horse	SLC5A3 (FIS)	QMYLYIQEVADYLTPLVAALFLLAIFWKRCN
Horse	SLC5A1	QLFDYIQSITSYLGPPIAAVFLLAIFWKRTT
Horse	SLC5A2	QLFDYIQAVSSYLAPPVSAVFLALFVPRVN
Horse	SLC5A4	QLVHYIESISSYVGPPIAAVFLLAIFCKRVN
Horse	SLC5A9	QLFDYIQSVTSYLAAPPITALFLLAIFCKRVT
Horse	SLC5A10	QLFIYMQSVTSSLAPPVTAVFVLGIFWQRAN
Horse	SLC5A11	QLFIYIQSISSYLQPPVAVVFIMGCFWKRTN

Figure 3. SLC5A3 amino acid sequences alignments in the region of the FIS mutation. The amino acid affected by the mutation, indicated with an arrow, is conserved in all placental mammals now sequenced with high coverage (n = 11); the top panel shows alignments of this gene region in a selection of these mammals. The bottom panel shows alignments of this region with other members of the SLC5 gene family in the horse which show structural similarity; the amino acid affected by the mutation is conserved in all of these members.
doi:10.1371/journal.pgen.1002133.g003

response mechanisms have also been found in numerous tissues, and in particular are critical for lymphocyte development and function [17,18].

The mechanism by which the osmotic stress response is mediated in mammals is not completely understood, but involves a signaling cascade comprising Rho-type small G-proteins, p38 Mitogen-activated protein kinase (p38MAPK) and the transcription factor, Nuclear Factor of Activated T-cells 5 (NFAT5) [19]. NFAT5 directly stimulates the transcription of hyperosmolarity-responsive genes, of which *SLC5A3* is one. These act to counterbalance the effects of extracellular osmotic pressure by transporting small organic osmolytes, such as myo-inositol, into the cell, thereby maintaining isotonicity with respect to extra-cellular conditions [20]. NFAT5 is the only known transcriptional activator of hyperosmolarity response genes, and was shown to be essential for normal lymphocyte proliferation and adaptive immunity [18]. Targeted knockout of *NFAT5* in mice results in late gestational lethality whereas partial loss of function leads to defects in adaptive immunity and a substantially reduced spleen and thymus [18]. Furthermore, transgenic studies identify loss of T-cell mediated immunity as the prime deficiency ensuing from aberrant NFAT5 activity [21,22]. Similarly FIS-affected foals have markedly reduced thymus and spleen with a lack of germinal centers [1,23]. However, FIS disease immunopathology indicates that FIS foals have apparently normal circulating T-cell numbers with only peripheral blood B-lymphocyte numbers significantly depleted; currently, there are no data available indicating NFAT5 activity in specific B-lymphocyte functions.

Studies are now required to demonstrate the functional differences between the Pro446 and Leu446 forms of the protein. This will be achieved by introducing this mutation into transgenic mice and assessing transport function. Further investigation into the physiological consequences of this mutation will then also be possible. In particular, it will be important to identify how the mutation leads to profound anemia and B-lymphopenia whilst neutrophils and T-cell numbers (including CD4/CD8 ratio) and function (responses to mitogens PHA and Con A) appear normal [4]. Importantly, it must be investigated whether there is a defect in T-cell function that is currently undetected or whether the antigen presenting function of B-cells is so suppressed that the T-cells cannot respond. In addition, it must be noted that the lymphoid organs in FIS foals have depleted thymus tissue and poor germinal centre development in spleen and lymph nodes, which suggests that there may be some unidentified T-cell dysfunction. Alternatively, any T-cell defects could be due to severe inflammatory responses in these very sick foals.

SLC5A3 is not associated with any described mammalian disease, although a role in the pathogenicity of Down Syndrome is suggested [24]. The effect of loss of SLC5A3 activity has not been comprehensively studied, however *SLC5A3* knockout mice die shortly after birth due to hypoventilation [25], probably due to failure of the peripheral nervous system [26]; similarly FIS-affected foals have peripheral ganglionopathy [1].

There is relatively little literature on SLC5A3 function in hemopoietic or immunological tissues, in any species, although a

role for osmotic control in developing cells is likely. Due to uncertainty regarding tissue distribution and function of SLC5A3, it cannot be assumed that the profound anemia and severe loss of circulating B-lymphocytes in FIS is directly due to a functional change in SLC5A3 expression or function; formal proof that this is the case will entail functional studies. Whilst there is no doubt that the mutation in this gene is predictive of carrier or disease status, the mechanism by which this amino acid change could lead to the two described pathologies is speculative. It is, of course possible that the mutation site is close to another, as yet unidentified, mutation that is ultimately responsible for the severe hematological and immunological changes in homozygotes. However, all coding sequence within the critical region has been fully investigated and this is the only variant that segregates with the disease. We hypothesize that the phenotype seen in FIS-affected foals is either a result of partial loss or subtle alteration in SLC5A3 activity that has deleterious effects on B-lymphocyte and erythroid development but cannot discount the involvement of other genetic variants in the critical region. Whichever is the case, further analysis of FIS is justified as this genetically determined combination of immune phenotypes has not previously been reported in any other species.

Materials and Methods

Ethics Statement

Procedures were limited to the collection of blood by jugular venipuncture or hairs pulled from the mane or tail. Blood samples were taken as veterinary diagnostic procedures as all study animals were equine patients presenting with clinical signs suggestive of FIS or were healthy related or unrelated animals that were blood tested for anemia and/or B-lymphocyte deficiency.

Study Population and Diagnostic Procedures

Many of the Fell and Dales ponies used in this study have been described previously [6,7]. Study animals were all equine patients presenting with clinical signs suggestive of FIS or were healthy related or unrelated Fell or Dales ponies that were blood tested for anemia and/or B-lymphocyte deficiency. Several FIS foals presented subsequent to euthanasia. Pedigree information was available for many of the Fell ponies (Figure S1), and these samples ($n = 41$) were used in the linkage and homozygosity mapping analysis. An additional ten samples were added to these for the association study; these were isolated samples for which no pedigree information and/or samples from immediate family were available. Any adult Fell pony was eligible as a control for the association study.

FIS diagnosis was based on breed, age of animal (4–8 weeks at presentation), and profound anemia with no other predisposing cause, and was confirmed on pathology. Specifically, this indicated severely reduced numbers (or absence) of germinal centers in spleen and regional lymph nodes. B-lymphocyte deficiency was also used for FIS diagnosis. Many accompanying clinical signs were also reported, primarily related to opportunistic infections, but these were not considered diagnostic alone.

Samples

Blood samples were collected in EDTA collection tubes from all of the Fell pony individuals indicated in Figure S1, and from a Dales foal and its parents [6]. Genomic DNA was isolated from the samples using a NucleonTM BACC Genomic DNA Extraction Kit.

Microsatellite Markers

A panel of 228 markers, distributed as evenly as possible over the equine genome and described in Table S1, was used. Two

further markers, TKY1155 and TKY2012, which were located in the implicated region, were subsequently genotyped.

The genome scan was performed in multiplexes of three markers. Four PCR reactions, each utilising a different fluorescent dye, were pooled together post-PCR to form a panel of 12 markers for analysis. An 18 bp tail (5'-TGACCGGCAGCAAAATTG-3') was added to the 5' end of the forward primer and a complementary fluorescent labelling primer was included in the PCR reaction as a means of making the reactions more efficient and to reduce costs [27]. Amplification was performed in 6 μ l volumes, using 2.5 pmol of reverse, 1 pmol of tailed-forward, 5 pmol of the labelled universal primer (either 6-FAM, VIC, NED, or PET), 20 ng genomic DNA, 0.75 unit AmpliTaq Gold (Applied Biosystems), 1 \times GeneAmp PCR buffer II (Applied Biosystems), 1.5 mM MgCl₂, and 200 μ M each dNTP. After denaturation at 94°C for 10 min, a 30-cycle PCR of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by 8 cycles of 94°C for 1 min, 50°C for 1 min, and 72°C for 1 min was performed, followed by a final extension at 72°C for 30 min. Genotyping analysis was performed on an ABI3100 (Applied Biosystems) according to the manufacturer's instructions. Genotyping data was analysed with GeneMapper version 4.0 (Applied Biosystems); alleles were assigned to pre-defined bins and automatically given an appropriate integer value. Mendelian inheritance was checked.

Linkage and Homozygosity Mapping

We used 41 ponies (14 FIS-affected, 17 obligate carriers, 10 adults of unknown carrier status) for which pedigree information and DNA was available, in the linkage analysis (Figure S1).

A parametric linkage analysis was carried out using SUPERLINK v.1.5 [28] assuming an autosomal recessive mode of inheritance. The disease allele frequency was estimated at 0.1, with 100% penetrance.

Pearson's χ^2 test of independence was used to identify markers where homozygosity varied significantly between the cases and controls. An $A \times 2$ (where $A =$ number of alleles at a given locus) contingency table with $A-1$ degrees of freedom was used. Expected and observed heterozygosity values were computed for cases and controls for all markers exhibiting a positive LOD score, using ARLEQUIN [29]. Statistical significance was assessed by calculating the one-tailed probability of the chi squared distribution.

Genome-Wide Association Mapping

SNP genotyping on 51 genomic DNA samples was performed using standard manufacturer's protocols by Cambridge Genomic Services (University of Cambridge, UK). The Illumina EquineSNP50 Infinium BeadChip, which contains 54,602 validated SNPs, was used; information on this array is available at http://www.illumina.com/documents/products/datasheets/datasheet_equine_snp50.pdf. Quality control and genotype calling were performed using GenomeStudio v.2009.2 (Illumina Inc.). Samples with a call rate $<95\%$ were discarded ($n = 2$). We performed a basic case-control association analysis on the remaining 49 samples (18 affected and 31 controls). Analysis was performed with the software package PLINK [30]. SNPs with low minor allele frequency (<0.02) or genotyping rate ($<90\%$) were excluded; this left 42,536 SNPs for analysis. The presence of population stratification was assessed using multi-dimensional scaling (Figure S2) and quantile-quantile plots were drawn to confirm that there was no over-inflation of the test statistics (Figure S3).

Fine Mapping

A total of 62 polymorphic SNPs were studied in 13 affected individuals. A subset of these helped to delineate recombination breakpoints and these are identified in Figure 2. Information

regarding these SNPs can be found at http://www.broadinstitute.org/ftp/distribution/horse_snp_release/v2/eqcab2.0_chr26_snps.xls.

PCR amplification of the target sequence containing each informative SNP was performed in 12 μ l volumes containing 20 ng genomic DNA, 0.75 unit AmpliTaq Gold, 1 \times GeneAmp PCR buffer II, 1.5 mM MgCl₂, 200 μ M each dNTP, 10 pmol of reverse and of tailed-forward primer. A PCR program of 94°C for 10 min, followed by 30 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 2 min, and then an extension of 72°C for 10 min was used. The PCR products were purified (MultiScreen PCR₉₆ filter plates; Millipore) before sequencing in a 6 μ l volume using 0.5 μ l of 5 \times BigDye Terminator v3.1 (Applied Biosystems), 5–20 ng PCR template, 1 μ l of 1 \times BigDye sequencing buffer and 3.2 pmol universal sequencing primer (Sigma-Aldrich). Templates >500 bp were also sequenced in the reverse direction. Sequencing was performed using cycle sequencing: 96°C for 0.5 min, 44 cycles of 92°C for 4 s, 58°C for 4 s and 72°C for 1.5 min. Purification was performed by isopropanol precipitation followed by sequencing on an ABI3100 according to the manufacturer's instructions. Sequences were viewed using STADEN [31].

Re-Sequencing of Candidate Region

This was performed at the Centre for Genomic Research (University of Liverpool, UK). A region of 3 Mb (ECA26: 28,942,655 – 31,942,655 Mb) was selected for re-sequencing which encompassed the critical region. Custom tiling 385 k NimbleGen Sequence Capture arrays (<http://www.454.com/products-solutions/experimental-design-options/nimblegen-sequence-capture.asp>) which covered 92.9% of the target were designed from the horse reference sequence using standard repeat-masking algorithms. Five individuals were selected for re-sequencing consisting of one affected pony (A13 in Figure 2), its parents, one obligate carrier selected for maximal homozygosity over the region and one individual apparently homozygous wild-type.

Sequencing was performed using GS FLX Titanium Series chemistry and assembled using Roche Newbler software v2.0.00. An average 34-fold read depth was obtained. Sequence from each of the five sequenced animals was aligned to the EquCab2 reference sequence using the Artemis Comparison Tool (ACT) [32] to identify possible rearrangements or insertion/deletions (Figure S4).

Identifying Candidate Mutations for FIS

MySQL (Oracle Corporation) was used to interrogate the data. The critical region was narrowed using heterozygous variants in the affected foal and Sanger sequencing subsequently verified these. The narrowed critical region was then interrogated for variants that segregated as expected for a recessive mutation; putative causal variants were confirmed or disproved using Sanger sequencing.

Supporting Information

Figure S1 The Fell pony pedigrees used for linkage analysis. Affected (FIS) individuals are shown shaded in black and obligate carriers are indicated with a dot. Individuals that are neither affected or obligate carriers are shown un-shaded. Individuals also coloured yellow were genotyped and used for linkage and homozygosity mapping. Double lines indicate consanguinity. (TIF)

References

- Scholes SF, Holliman A, May PD, Holmes MA (1998) A syndrome of anemia, immunodeficiency and peripheral ganglionopathy in Fell pony foals. *Vet Rec* 142(6): 128–34.

Figure S2 Multidimensional scaling plot of first two components. Multidimensional scaling analysis illustrating the first two components of Identity-By-State similarity for all FIS-affected (red squares) and controls (blue circles) used in the genome-wide association analysis. A permutation test (10,000 permutations) for between-group IBS differences showed that there was no significant difference between the affected and controls ($P=0.553$). (TIF)

Figure S3 Observed versus expected $\log_{10}(P)$. A Q-Q plot showing the distribution of expected versus observed $-\log_{10}P$ for the basic association test with no adjustment. The pink diagonal shows the values expected under the null hypothesis. The observed $-\log_{10}P$ values match the expected values along the major portion of the graph and deviate towards the end illustrating the small number of true associations. The plot indicates minimal population stratification and therefore no corrections were subsequently made to the data. (TIF)

Figure S4 Alignment of re-sequenced region from FIS pony to the EquCab2 reference sequence. Sequence from the FIS-affected pony was aligned to the EquCab2 reference sequence using the Artemis Comparison Tool (ACT) [32] to identify possible rearrangements, duplications or insertion/deletions. Red indicates alignment of the sequence. White regions indicate missing sequence caused either by missing probe design, or small tandem repeats which result in reads stacking on top of each other and causing a break in the contig. Blue lines indicate inverted repeats scattered throughout the sequence. Black lines represent blocks of synteny used for sequence comparison and do not represent sequence variation. These alignments provide no evidence for significant rearrangement, duplication or insertion/deletion within the sequences; there is an excellent match between the FIS pony and the reference sequence. (TIF)

Table S1 Panel of 228 microsatellite markers used in linkage and homozygosity mapping. (DOC)

Table S2 Linkage analysis and homozygosity mapping of FIS using a genome-wide microsatellite set. (DOC)

Acknowledgments

We thank the Fell and Dales Pony Societies and their members for their support and providing invaluable samples. We thank the veterinary surgeons and histologists who have collected and analyzed samples, particularly N. Flindall and T. Blunden. We thank Cambridge Genomic Services at the University of Cambridge and C. Sibbons for performing the genotyping. Thanks to the Centre for Genomic Research at the University of Liverpool for performing the sequence capture and re-sequencing, and L. Downs and D. Adelson for their assistance with analysis. Thanks to H. Browne and D. Grafham of the Sanger Institute for their help with ACT.

Author Contributions

Conceived and designed the experiments: SDC WEO JES. Performed the experiments: LYF-C IG JES. Analyzed the data: LYF-C IG JES. Contributed reagents/materials/analysis tools: NH DCK PDFM. Wrote the paper: JES.

- Thomas GW, Bell SC, Phythian C, Taylor P, Knottenbelt DC, et al. (2003) Aid to the antemortem diagnosis of Fell pony foal syndrome by the analysis of B lymphocytes. *Vet Rec* 152(20): 618–21.

3. Thomas GW, Bell SC, Carter SD (2005) Immunoglobulin and peripheral B-lymphocyte concentrations in Fell pony foal syndrome. *Equine Vet J* 37(1): 48–52.
4. Bell SC, Savidge C, Taylor P, Knottenbelt DC, Carter SD (2001) An immunodeficiency in Fell ponies: a preliminary study into cellular responses. *Equine Vet J* 33(7): 687–92.
5. Dixon JB, Savage M, Wattret A, Taylor P, Ross G, et al. (2000) Discriminant and multiple regression analysis of anemia and opportunistic infection in Fell pony foals. *Vet Clin Pathol* 29(3): 84–86.
6. Fox-Clipsham L, Swinburne JE, Papoula-Pereira RI, Blunden AS, Malalana F, et al. (2009) Immunodeficiency/anemia syndrome in a Dales pony. *Vet Rec* 165(10): 289–90.
7. Thomas GW (2003) Immunodeficiency in Fell Ponies. PhD thesis:University of Liverpool.
8. Shin EK, Perryman LE, Meek K (1997) A kinase-negative mutation of DNA-PK(CS) in equine SCID results in defective coding and signal joint formation. *J Immunol* 158(8): 3565–9.
9. Wiler R, Leber R, Moore BB, VanDyk LF, Perryman LE, et al. (1995) Equine severe combined immunodeficiency: a defect in V(D)J recombination and DNA-dependent protein kinase activity. *Proc Natl Acad Sci U S A* 92(25): 11485–9.
10. McGuire TC, Banks KL, Davis WC (1976) Alterations of the thymus and other lymphoid tissue in young horses with combined immunodeficiency. *Am J Pathol* 84(1): 39–54.
11. Perryman LE (2000) Primary immunodeficiencies of horses. *Vet Clin North Am Equine Pract* 16(1): 105–16. vii.
12. Faham S, Watanabe A, Besserer GM, Cascio D, Specht A, et al. (2008) The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na⁺/sugar symport. *Science* 321(5890): 810–4.
13. Watanabe A, Choe S, Chaptal V, Rosenberg JM, Wright EM, et al. (2010) The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature* 468(7326): 988–91.
14. Brandl CJ, Deber CM (1986) Hypothesis about the function of membrane-embedded proline residues in transport proteins. *Proc Natl Acad Sci U S A* 83(4): 917–21.
15. Vilsen B, Andersen JP, Clarke DM, MacLennan DH (1989) Functional consequences of proline mutations in the cytoplasmic and transmembrane sectors of the Ca²⁺(+)-ATPase of sarcoplasmic reticulum. *J Biol Chem* 264(35): 21024–30.
16. Haussinger D (1996) The role of cellular hydration in the regulation of cell function. *Biochem J* 313 (Pt 3): 697–710.
17. Kino T, Takatori H, Manoli I, Wang Y, Tiulpakov A, et al. (2009) Brx mediates the response of lymphocytes to osmotic stress through the activation of NFAT5. *Sci Signal* 2(57): ra5.
18. Go WY, Liu X, Roti MA, Liu F, Ho SN (2004) NFAT5/TonEBP mutant mice define osmotic stress as a critical feature of the lymphoid microenvironment. *Proc Natl Acad Sci U S A* 101(29): 10673–8.
19. Kino T, Segars JH, Chrousos GP (2010) Brx, a link between osmotic stress, inflammation and organ physiology/pathophysiology. *Expert Rev Endocrinol Metab* 5(4): 603–614.
20. Burg MB, Ferraris JD, Dmitrieva NI (2007) Cellular response to hyperosmotic stresses. *Physiol Rev* 87(4): 1441–74.
21. Trama J, Go WY, Ho SN (2002) The osmoprotective function of the NFAT5 transcription factor in T cell development and activation. *J Immunol* 169(10): 5477–88.
22. Morancho B, Minguillon J, Molkenin JD, Lopez-Rodriguez C, Aramburu J (2008) Analysis of the transcriptional activity of endogenous NFAT5 in primary cells using transgenic NFAT-luciferase reporter mice. *BMC Mol Biol* 9: 13.
23. Richards AJ, Kelly DF, Knottenbelt DC, Cheeseman MT, Dixon JB (2000) Anemia, diarrhoea and opportunistic infections in Fell ponies. *Equine Vet J* 32(5): 386–91.
24. Berry GT, Wang ZJ, Dreha SF, Finucane BM, Zimmerman RA (1999) In vivo brain myo-inositol levels in children with Down syndrome. *J Pediatr* 135(1): 94–7.
25. Berry GT, Wu S, Buccafusca R, Ren J, Gonzales LW, et al. (2003) Loss of murine Na⁺/myo-inositol cotransporter leads to brain myo-inositol depletion and central apnea. *J Biol Chem* 278(20): 18297–302.
26. Chau JF, Lee MK, Law JW, Chung SK, Chung SS (2005) Sodium/myo-inositol cotransporter-1 is essential for the development and function of the peripheral nerves. *FASEB J* 19(13): 1887–9.
27. Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 18(2): 233–4.
28. Silberstein M, Tzemach A, Dovgolevsky N, Fishelson M, Schuster A, et al. (2006) Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *Am J Hum Genet* 78(6): 922–35.
29. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 1: 47–50.
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3): 559–75.
31. Staden R, Beal KF, Bonfield JK (2000) The Staden package, 1998. *Methods Mol Biol* 132: 115–30.
32. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* 21(16): 3422–3.

Papers

Population screening of endangered horse breeds for the foal immunodeficiency syndrome mutation

L. Y. Fox-Clipsham, E. E. Brown, S. D. Carter, J. E. Swinburne

The Fell and Dales are UK pony breeds that have small populations and may be at risk from in-breeding and inherited diseases. Foal immunodeficiency syndrome (FIS) is a lethal inherited disease caused by the recessive mutation of a single gene, which affects both Fell and Dales ponies and potentially other breeds that have interbred with either of these. FIS, previously known as Fell pony syndrome, is characterised by progressive anaemia and severe B lymphocyte deficiency. The identification of the causal mutation for this disease led to the recent development of a DNA-based carrier test. In this study, the authors used this test to estimate the prevalence of the FIS mutation in the Fell and Dales populations, revealing that approximately 18 per cent of adult Dales ponies and 38 per cent of adult Fell ponies are carriers of the FIS defect. In addition, a study of five potential at-risk breeds was conducted to assess the transfer of the FIS defect into these populations. Of the 192 coloured ponies tested, two were confirmed as FIS carriers: No carriers were found among 210 Clydesdales, 208 Exmoor ponies, 161 Welsh section D, 49 part-bred Welsh section D and 183 Highland ponies.

THE Fell and Dales are related pony breeds that are native to the British Isles and are registered by the Rare Breeds Survival Trust (RBST 2011), because of their low numbers (Fell pony <1500 registered breeding mares and Dales pony <500 registered breeding mares). Both breeds are descended from relatively few animals and have experienced near extinction. The Fell pony experienced a genetic bottleneck, resulting in a loss of genetic diversity shortly after the Second World War, when the coming of tractors resulted in the slaughter of many redundant ponies (Thomas 2003). It is therefore highly probable that this recent genetic bottleneck, coupled with overuse of prominent sires led to the emergence of a fatal inherited defect, foal immunodeficiency syndrome (FIS). FIS was first reported in Fell ponies in 1998, at a time when many cases were identified, and in 2009, the first case was reported in a Dales foal (Fox-Clipsham and others 2009). FIS-affected foals are normal at birth but within the first few weeks of life, usually before four weeks of age, begin to lose condition and show signs of this disease (Fig 1). Most commonly, foals are present with weight loss and a poor demeanour and clinical signs symptomatic of multiple infections, including diarrhoea, weight loss, nasal discharge and coughing (Scholes and others 1998). Despite extensive treatment, these infections become unresponsive and persistent, resulting in the loss of foals before 16 weeks of age. The underlying pathology causing these clinical signs has been shown to be anaemia (Dixon and others

2000, Richards and others 2000), which is unrelated to blood loss or haemolysis, and a severe B lymphocyte deficiency (Thomas and others 2003) with reduced antibody production (Thomas and others 2005). FIS results in 100 per cent mortality. Pedigree analysis, incorporating the knowledge of affected animals, strongly suggests an autosomal recessive mode of inheritance (Thomas 2003). Furthermore, pedigree analysis has identified the likely founder as a stallion from the 1950s that features in both the maternal and paternal ancestry of all FIS-affected foals (Fox-Clipsham 2011). Carriers are phenotypically normal, with offspring of carriers having a 50 per cent risk of inheriting the defect. Carrier-carrier matings are the only combination that can give rise to affected individuals, with each carrier-carrier mating having a 25 per cent chance of producing affected offspring (Table 1), by inheriting one copy of the defect from each parent. Until recently, carriers could only be confirmed by the appearance of affected offspring. Recently, however, the genetic defect that causes FIS has been genetically mapped and characterised (Fox-Clipsham and others 2011). As a result, in February 2010, a carrier test for FIS was launched by the Animal Health Trust (AHT), enabling the carrier status of the animal to be identified from a hair sample (AHT 2011). Thus, it is now possible to make informed and confident breeding decisions, avoiding carrier-carrier matings and the fear of producing affected offspring. During the period of February to December 2010, 888 samples were submitted for FIS screening, either to identify breeding carriers or to confirm the diagnosis of suspected FIS-affected foals.

Potentially, FIS could affect other equine breeds that have Fell or Dales ancestry or have interbred with the Fell and Dales population over recent years. Indeed, it is well recognised that, over recent decades, both the Fell and Dales have interbred with other breeds, particularly sturdy native breeds. Discussions with breeders and societies revealed that Clydesdales, Exmoor, Highland, Welsh section D, part-bred Welsh section D and coloured horses and ponies were most likely to have been used for interbreeding with potential FIS carriers, because of their similar type and geographical location. Thus, a study was designed, using the DNA-based carrier test, to identify and enumerate the FIS carriers in populations of Fell and Dales ponies and also in the

Veterinary Record

doi: 10.1136/vr.100235

L. Y. Fox-Clipsham, BSc, PhD,

E. E. Brown, BSc,

J. E. Swinburne, BSc, PhD,

Centre for Preventive Medicine, Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk, CB8 7UU, UK

S. D. Carter, BSc, PhD, FRCPath, Department of Infection Biology, School of Veterinary Science, University of Liverpool, Liverpool, L69 7ZJ, UK

E-mail for correspondence:

laurayclipsham@hotmail.co.uk

Provenance: not commissioned; externally peer reviewed

Accepted September 19, 2011



FIG 1: Fell pony foal showing typical of FIS: poor demeanour, dull coat, diarrhoea, weight loss and nasal discharge

other 'potential at-risk' breeds. Data generated by this study would be useful to breed societies and individual horse breeders and owners to eliminate this fatal disease from equine populations.

The first objective of this study was to provide an estimation of the carrier prevalence by performing an anonymous and random screen of the Fell and Dales population. In addition, results from the testing performed at the AHT since the launch in February 2010 are presented. Secondly, a screen of those breeds that are known to have interbred with the Fell and Dales was performed to assess the spread of the FIS mutation.

Materials and methods

Animals

Two hundred and fourteen adult Fell pony hair samples and 87 adult Dales pony hair samples, which were among those sent to the AHT for DNA (parentage) profiling between 2000 and 2010 and thereafter archived were randomly and anonymously selected for FIS screening. All of these samples were from the UK populations only. In addition, 210 Clydesdale, 208 Exmoor pony, 161 Welsh section D, 49 part-bred Welsh section D, 92 Highland pony and 90 coloured horse and pony hair samples, submitted between 2009 and 2010 were also selected from this archive. Permission was obtained from each respective breed society for the analysis of these samples. An additional 91 Highland pony and 102 coloured pony hair samples submitted by breeders for the specific purpose of this population screen were also analysed.

Genotyping of individuals

For each sample, DNA was extracted from six equine hair roots using a 5 mg/ml proteinase K extraction. The extract was heated to 60°C for 45 minutes and then 95°C for 15 minutes. DNA primers, with a universal 18 base pair tail (5'-TGACCGGCAGCAAAATTG-3') added to the 5' end of the forward primer, (forward primer 5' CTCATGATTGTGGGGAGGATA-3'; reverse primer 5'-ATCAGGTTGGTCACATTCTGG-3') were used in a polymerase chain reaction (PCR) with 1.5 mM MgCl₂ to amplify the affected region from the DNA sample. The target region is 282 nucleotides in length. The following PCR protocol was used to amplify the target region: 94°C for 10 minutes, followed by 30 cycles of 94°C for one minute, 58°C for one minute and 72°C for two minutes and then 72°C for 10 minutes. The PCR products were purified (MultiScreen PCR₉₆ filter plates; Millipore) before sequencing in a 6 µl volume using 0.5 µl of 5x BigDye Terminator v3.1 (Applied Biosystems), 5 to 20 ng PCR template, 1 µl of 1x BigDye sequencing buffer and 3.2 pmol universal

TABLE 1: Expected outcomes (% of foals) from the matings of genotypically normal ponies and carriers of the FIS mutation

Matings of ponies	Expected outcomes (%)		
	Normal foals	Carrier foals	Affected (FIS) foals
Normal x normal	100	0	0
Normal x carrier	50	50	0
Carrier x carrier	25	50	25

FIS Foal immunodeficiency syndrome

sequencing primer. Sequencing was performed using cycle sequencing: 96°C for 0.5 minutes, 44 cycles of 92°C for four seconds, 58°C for four seconds and 72°C for 1.5 minutes. Purification was performed by isopropanol precipitation followed by sequencing on an ABI 3100 automated sequencer according to the manufacturer's instructions. The FIS genetic defect is a single nucleotide polymorphism at nucleotide position 134 in the amplicon. Thus, an animal that is homozygous for the wild-type allele 'C' is clear of the FIS mutation, whereas an affected animal is homozygous for the mutant 'T' allele and a carrier is heterozygous 'C/T'. Examples of results generated for a 'clear' (non-carrier), 'carrier' and 'affected' animal are shown in Fig 2.

Results

Using the DNA-based carrier test, Fell and Dales pony populations were anonymously and randomly screened to provide an estimation of the carrier frequency for 2009/2010. Data from this analysis show that approximately 40 per cent of adult UK Fell ponies and 20 per cent of adult UK Dales ponies carried the FIS defect (Table 2).

To assess the potential spread of the FIS mutation, five additional breeds were screened using the DNA-based carrier test. Of the five at-risk breeds tested, no FIS carriers were identified among the Clydesdale, Exmoor, Highland or Welsh section D samples. However, of the 192 coloured horse and pony samples that were screened, two were confirmed as carriers of FIS. These data confirm that there has been some transfer of the mutation into the coloured population (Table 2). Both samples that were confirmed as carriers had been submitted specifically for the purpose of this study. After discussions with the owners of these ponies, it was established that one was known to have Fell ancestry, while the ancestry of the other pony was unknown.

In addition to the random screening of archive samples, data from those samples submitted for FIS testing have also been analysed. From the test launch in February 2010 until December 2010, a total of 888 samples were submitted to the AHT for FIS screening. Of these, 702 were from Fell ponies and the remaining 186 were from Dales ponies (Table 3). Samples submitted for FIS screening were received from the UK, mainland Europe, Canada and the USA. Data from this testing show that, of submitted samples, 11 per cent adult Dales ponies and 48 per cent adult Fell ponies carry the FIS defect. This test has also been used to confirm the diagnosis of suspected FIS foals and has successfully diagnosed FIS in 14 foals, including one Dales foal.

Discussion

Due to the recent identification of the causal mutation for FIS, the authors have been able to perform the first population screen of the UK adult Fell and Dales pony populations. This has provided an estimate of the number of carriers of FIS. The DNA test has also enabled a population screen of other equine breeds to assess the spread of the FIS defect. In 2001, a dramatic decline (15 to 25 per cent) was reported in the number of foals being registered by the Fell Pony Society, and this was thought to be mainly due to the loss of foals due to FIS (Bell and others 2001). These data suggested that approximately 40 to 50 per cent of adult Fell ponies carried the FIS defect in 2001.

In the current investigation, a selection of samples submitted to the AHT (submitted from 2000 to 2010) for DNA profiling were screened for FIS, identifying 18.40 per cent (16/87) of the Dales ponies and 38.32 per cent (82/214) of the Fell ponies as FIS carriers (Table 2). It should be noted that DNA (parentage) profiling is a requirement for licensing both Dales and Fell stallions and that on average twice as many stallions are submitted for DNA profiling compared with mares. It is therefore likely that this population screen is biased towards the

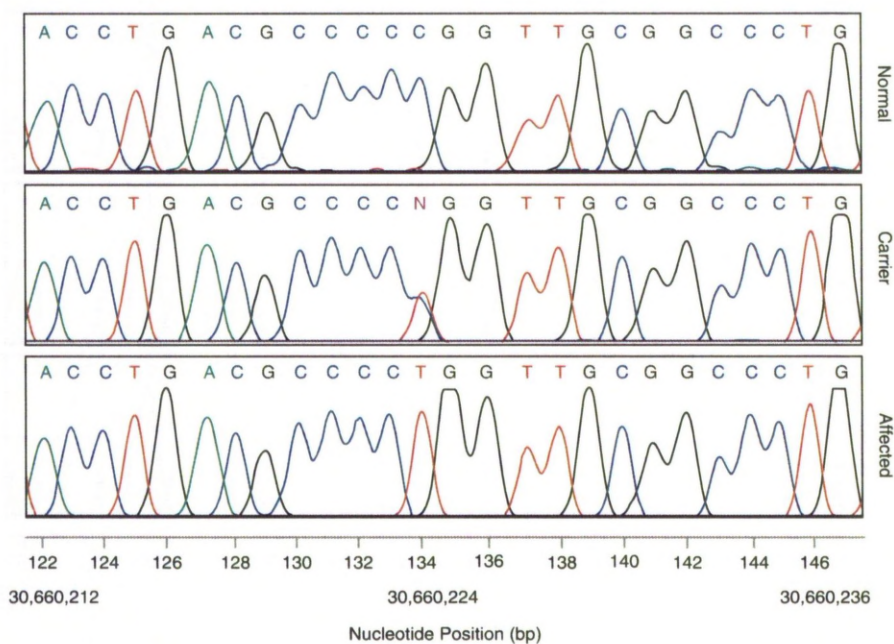


FIG 2: Sequencing traces obtained from a normal (homozygous wild-type), FIS carrier (heterozygous for mutant allele) and FIS-affected (homozygous for the mutant allele) pony. The mutation is at nucleotide position 30,660,224 bp on chromosome 26 (position 134 in the amplicon; grey bar)

TABLE 2: Results of the population screen using the FIS DNA-based carrier test in the Fell and Dales pony and five breeds which are considered at risk from the spread of the FIS defect

Breed	Number of non-carriers	Number of carriers	Total
Clydesdale	210 (100%)	0 (0%)	210
Coloured	190 (98.96%)	2 (1.04%)	192
Dales	71 (81.61%)	16 (18.39%)	87
Exmoor	208 (100%)	0 (0%)	208
Fell	132 (61.48%)	82 (38.52%)	214
Highland	183 (100%)	0 (0%)	183
Part-bred Welsh section D	49 (100%)	0 (0%)	49
Welsh section D	161 (100%)	0 (0%)	161

TABLE 3: Results from samples submitted to the FIS screening test which was launched in February 2010 by the Animal Health Trust to identify carriers and confirm the diagnosis of suspected FIS-affected foals

Samples	Number of non-carriers	Number of carriers	Number of affected animals	Total
Dales pony				
Carrier screening	165 (89.19%)	20 (10.81%)	0 (0%)	185
Suspected FIS foals	0 (0%)	0 (0%)	1 (100%)	1
Fell pony				
Carrier screening	356 (51.97%)	329 (48.03%)	0 (0%)	685
Suspected FIS foals	2 (11.76%)	2 (11.76%)	13 (76.47%)	17
FIS Foal immunodeficiency syndrome				

stallion population; the results from the Fell pony screen have estimated the carrier prevalence as ~40 per cent, which is less than that estimated for 2001. This drop in carrier numbers is probably due to the Fell breeders and owners having prior knowledge of breeding outcomes from previous carrier matings and being selective in their breeding programmes to avoid FIS foals being produced.

The data from the random Fell pony and Dales pony testing indicated a 40 and 20 per cent carrier rate, respectively, in adult ponies. However, when compared with samples submitted specifically for FIS testing once the test was publicised there were, not surprisingly, different positivity rates (Table 3). The Fell pony samples submitted for FIS testing had a 48 per cent carrier rate, a slight increase compared with the random Fell pony samples tested. Meanwhile, the Dales pony

samples submitted for FIS testing showed a reduction (10.8 per cent compared with 18.4 per cent) in FIS carrier rates. This discrepancy is almost certainly due to the choice of animals selected for FIS testing but does not hide the important fact that there are high numbers of carriers in both of these equine populations. The FIS test is not only used to identify carriers, but it is now also used to confirm the diagnosis of FIS-affected foals. Samples from suspect FIS foals are processed at the AHT within three working days. This not only prevents the extended suffering of affected foals (for which diagnosis had been very difficult), but also provides breeders and veterinarians with a definitive diagnosis antemortem. Since the launch of the test, 18 foal samples have been submitted that were suspected FIS cases, including one Dales foal. Of these suspects, 14 were confirmed as FIS-affected foals, including one Dales foal. Of the remaining four foals, two survived, one later confirmed with tetanus and the other chronic diarrhoea. Of the two foals that died, one had suspected *Rhodococcus equi* infection and the other had no postmortem.

As part of this investigation, other equine breeds that were considered to be at

risk from the spread of the FIS defect were also examined. Five breeds were screened, including the Clydesdale, Highland pony, Exmoor pony, Welsh section D pony and coloured horses and ponies. Two FIS carriers were identified from the 192 coloured horse and pony samples screened, confirming that there has been a transfer of the FIS defect into this equine population. To prevent the further spread of this defect, breeders should be cautious in their breeding decisions. Owners of coloured horses and ponies with known Fell or Dales ancestry should consider screening their stock for carriers, enabling breeders to exclude carriers from breeding programmes and prevent the further spread of the FIS mutation among the coloured population. In this study, only registered animals were screened for the FIS mutation; however, it is quite likely that the FIS mutation will have spread into crossbred, unregistered animals and therefore particular caution should be taken by all breeders when cross breeding with Fell or Dales ponies, using only animals that have been proven clear of the FIS defect.

With DNA testing for FIS now available, Fell and Dales pony breeders who suspect that they have the FIS defect among their stock can test for carriers so that they can make informed breeding decisions. The main advantage of the FIS test is that carriers can be reliably identified so that breeders can confidently mate carriers of FIS with ponies that are clear of FIS, with no fear of producing an affected foal. Importantly, the authors consider that FIS carriers should not be excluded from Fell and Dales breeding programmes as this could reduce the gene pool, further endangering these breeds. Instead, Fell pony and Dales pony breeders and owners are recommended to avoid carrier-carrier matings in order to avoid FIS foals being produced, and, in the long term, gradually replace carriers with FIS clear animals. As a result, FIS carriers will gradually be phased out of breeding programmes. Elimination of a defect from a population requires careful management, but in time and with informed breeding decisions, the FIS mutation will gradually decline and be eradicated and the spread of the mutation into other breeds prevented.

Acknowledgements

The authors thank the Clydesdale Horse Society, the Coloured Horse and Pony Society, the Dales Pony Society, the Fell Pony Society, the Exmoor Pony Society, the Highland Pony Society and the Welsh Pony and Cob Society and their members for their support and for providing samples for this investigation. They thank the veterinary surgeons and individual breeders who have collected samples, particu-

larly Paul May at Townend Veterinary Practice, Derek Knottenbelt at the Equine Hospital, Leahurst, and Frame, Swift and Partners Veterinary Centre. They also thank the Equine DNA-typing Unit at the Animal Health Trust for performing the testing and providing us with data. This work was funded by The Horse Trust, who also supported LF-C.

References

- AHT (2011) Foal immunodeficiency syndrome. www.aht.org.uk/genetics_fis.html. Accessed July 4, 2011
- BELL, S. C., SAVIDGE, C., TAYLOR, P., KNOTTENBELT, D. C. & CARTER, S. D. (2001) An immunodeficiency in Fell ponies: a preliminary study into cellular responses. *Equine Veterinary Journal* **33**, 687-692
- DIXON, J. B., SAVAGE, M., WATTRET, A., TAYLOR, P., ROSS, G., CARTER, S. D., KELLY D.F., HAYWOOD, S. & OTHERS (2000) Discriminant and multiple regression analysis of anaemia and opportunistic infection in Fell pony foals. *Veterinary Clinical Pathology* **29**, 84-86
- FOX-CLIPSHAM, L., SWINBURNE, J. E., PAPOULA-PEREIRA, R. I., BLUNDEN, A. S., MALALANA, F., KNOTTENBELT, D. C. & CARTER, S. D. (2009) Immunodeficiency/anaemia syndrome in a Dales pony. *Veterinary Record* **165**, 289-290
- FOX-CLIPSHAM, L. Y. (2011) Foal immunodeficiency syndrome: identification of the causal mutation. In *Veterinary Pathology*. University of Liverpool
- FOX-CLIPSHAM, L. Y., CARTER, S. D., GOODHEAD, I., HALL, N., KNOTTENBELT, D. C., MAY, P. D., OLLIER, W. E. & SWINBURNE, J. E. (2011) Identification of a mutation associated with fatal Foal Immunodeficiency Syndrome in the Fell and Dales pony. *PLoS Genetics* **7**, e1002133
- RBST (2011) Watch list. www.rbst.org.uk/watch-list/equines. Accessed July 4, 2011
- RICHARDS, A. J., KELLY, D. E., KNOTTENBELT, D. C., CHEESEMAN, M. T. & DIXON, J. B. (2000) Anaemia, diarrhoea and opportunistic infections in Fell ponies. *Equine Veterinary Journal* **32**, 386-391
- SCHOLES, S. E., HOLLIMAN, A., MAY, P. D. & HOLMES, M. A. (1998) A syndrome of anaemia, immunodeficiency and peripheral ganglionopathy in Fell pony foals. *Veterinary Record* **142**, 128-134
- THOMAS, G. W. (2003) Immunodeficiency in fell ponies. In *Veterinary Pathology*. University of Liverpool
- THOMAS, G. W., BELL, S. C. & CARTER, S. D. (2005) Immunoglobulin and peripheral B-lymphocyte concentrations in Fell pony foal syndrome. *Equine Veterinary Journal* **37**, 48-52
- THOMAS, G. W., BELL, S. C., PHYTHIAN, C., TAYLOR, P., KNOTTENBELT, D. C. & CARTER, S. D. (2003) Aid to the antemortem diagnosis of Fell pony foal syndrome by the analysis of B lymphocytes. *Veterinary Record* **152**, 618-621