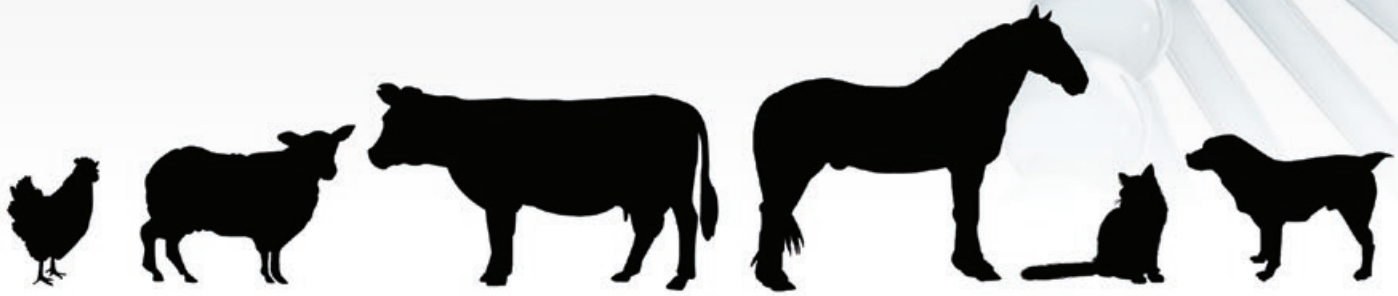
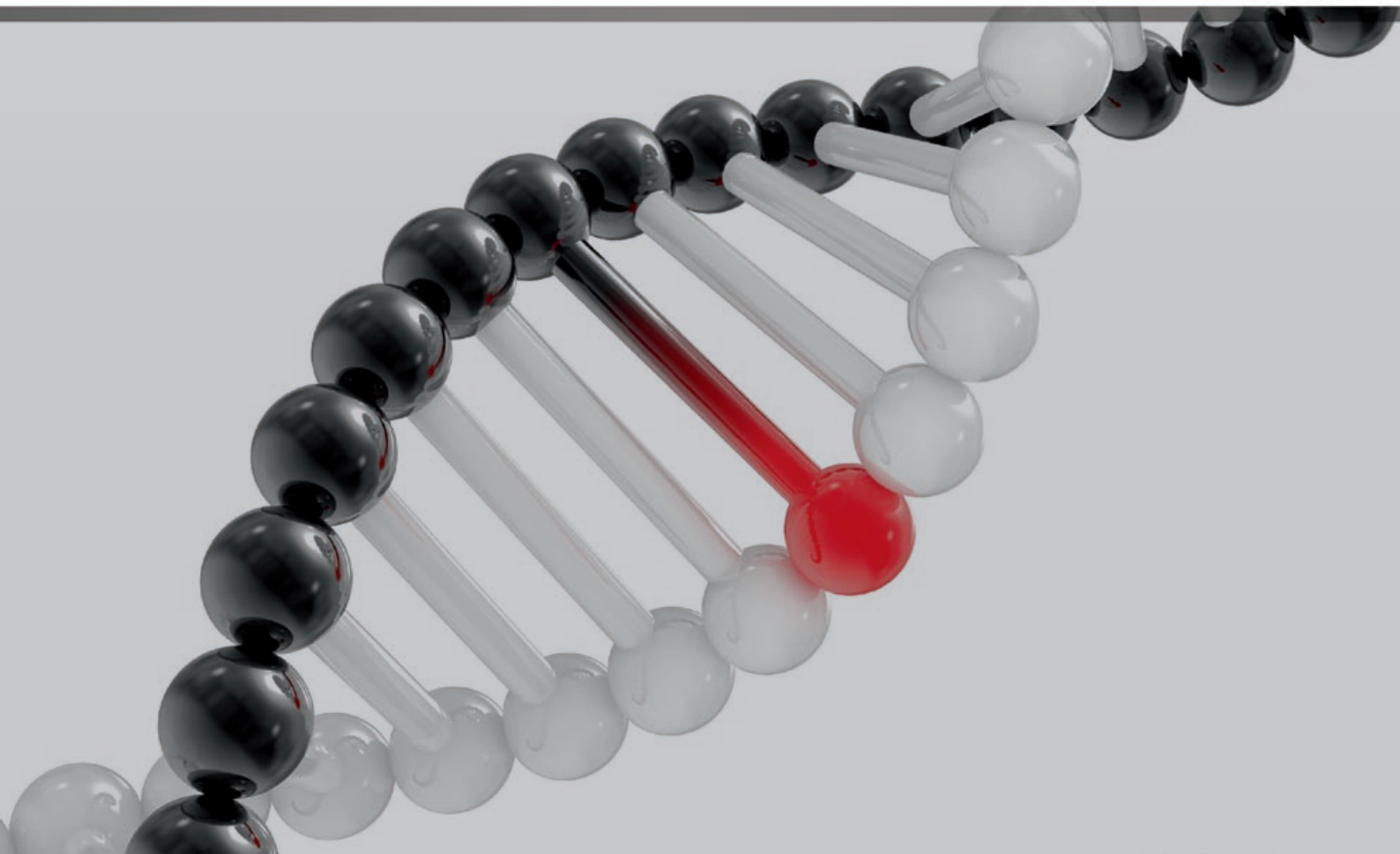
technical manual

RIDGENE

How to manage or remove inherited diseases from animal populations using “Homozygosity Mapping”





How to manage or remove inherited diseases from animal populations using 'Homozygosity Mapping'

RIDGENE
Manual Version 1.0

This manual has been produced as a result of work sponsored by:

Biosciences Knowledge Transfer Network (KTN)

Knowledge Transfer Network
Biosciences

Technology Strategy Board
Driving Innovation

Eblex

Quality Meat Scotland (QMS)

The work was carried out by the RIDGENE consortium:

The Roslin Institute, R(D)SVS, University of Edinburgh

Animal Health Trust

Scottish Agricultural College (now SRUC)

The content for the RIDGENE manual has been produced by John Woolliams, The Roslin Institute (R(D)SVS, University of Edinburgh) and edited by the Biosciences KTN.

Foreword

Welcome to the RIDGENE Manual, a technical guide for those looking to manage or remove recessive inherited diseases from animal populations.

Recessive inherited diseases can occur naturally in all animal species. When they appear, the effects can range from relatively minor through to severe consequences for the animal's health and welfare and even death. The ability to develop useful marker tests at relatively low cost to help manage any problems that arise, has therefore been high on the wish list for animal breeders for a long time.

In recent years there have been a number of examples where genetic tests have been successfully developed for animal species –such as CVM in Holstein dairy cattle may be familiar to many of you (and is described in more detail later in this manual if it is not). However the process of developing these tests has tended to be very costly, time consuming and has required a lot of data.

Recent developments in genomic technologies have opened the door to developing more cost effective approaches to generate useful genetic tests more quickly than in the past. Successful demonstration of a cost effective approach in animals was first published in 2008 (Charlier *et al*. 2008). This publication stimulated the joint initiative that has resulted in this manual.

The original aim was to develop a how-to manual to help breeders to: collect useful samples, assess if they have had a recessive disease problem, and to better understand the steps needed to go through to develop effective genetic markers. A series of real-world case studies were also conducted as part of the RIDGENE project, alongside the manual development, to demonstrate the process in practice – allowing the lessons learnt along the way to be highlighted. A number of discussion meetings were also held to inform interested members of progress and to have their input. The result is included in this booklet.

I would like to take this opportunity to thank the team that helped develop this manual, namely the other funders of the work (EBLEX and QMS), the academic consortium (particularly John Woolliams, from the Roslin Institute - who wrote most of this manual) and members of the Biosciences KTN team who helped edit this final version, and also the Biosciences KTN members that attended the dissemination events – everybody played a part in helping to deliver what I believe will be a very useful and valuable resource, I hope you think so too!

Huw Jones

Assistant Director (Animal Sector Lead), Biosciences KTN

Charlier et al., 2008, Nature Genetics Vol 40, no.4

Introduction

Recessive inherited diseases can occur naturally in all animal species. They arise when an animal inherits two copies of a defective version of a gene that results in a negative effect on the animal's physiology or development. When they appear, the effects can range from relatively minor through to severe consequences for the animal's health and welfare and even death.

In nearly all cases the defective version of a gene first appears due to an error in the process of DNA replication in a single animal. As the animal has only one copy at that time, the disease is not observed. If it happens to be an animal that is widely used for breeding, and their sons and/or daughters are also retained for breeding, it may not be long before two carriers are mated and the disease appears in some progeny. The main problem is that by the time it is first observed, there are many carriers in the population (particularly if previous cases have not been recorded), and numbers of disease cases will accumulate over time if not effectively managed.

Although errors in DNA replication may at first sound like an improbable event, in reality they are fairly common. Based on past research, it is estimated that every time DNA replicates in an animal an average of around 30 copying errors occur. The vast majority of these errors will have little or no effect on the developing animals, but on the odd occasion they can have significant effects – if positive, they can help drive evolution or selection, but if negative can result in disease. As a result, inherited diseases can arise at any time - and are not just old problems that have been circulating at low levels in a breed for many years. The ability to develop useful marker tests to help manage any problems that arise has therefore been high on the wish list for animal breeders for a long time.

This manual is intended to provide breeders with a how-to guide to achieve just that. It is focused around a process known as homozygosity mapping, which as a result of recent developments in genomic technologies, can provide the means to develop genetic tests in a cost effective manner. The work undertaken in order to develop this manual has included a number of demonstration projects with real world populations and real diseases, therefore the methodology described in the RIDGENE Manual has been tested and is a reliable practical guide not just a description of a theoretical method. These case studies are referenced in the text, particularly where lessons have been learned and a list of the species and types of diseases that have been investigated is included in Annex 2.

The manual aims to provide a structure for all those involved in the front line of recognising and responding to recessive diseases including farmers, breed societies, vets, extension organisations and researchers. With this aim in mind, the manual is split into a series of steps which can be examined separately or together.

The manual assumes a basic understanding of animal genomics and a glossary of terms is provided in Annex 1, with terms included in the glossary shown in bold blue within the manual. For more information on the basics of animal genomics, there are a number of useful textbooks and online resources available.

Contents

RIDGENE
Manual Version 1.0

Understanding recessive inherited disease

When is this manual relevant?

The incidence of all diseases is influenced to some degree by the **genes** inherited from parents. The degree can vary significantly, e.g. for foot and mouth, genetics has a minimal impact whereas for Foal Immunodeficiency Syndrome (FIS) in Fell Ponies it is the sole factor for whether or not the disease occurs. Other factors that influence disease are environmental, such as nutrition, management and degree of exposure to a pathogen or toxin.

Genetic influences on disease can be split into two classes, **simple** and **complex inheritance**:

> **Simple inheritance:**
> **where the genetic influence on the disease is controlled by just a single gene**
> Examples include: Spider Lamb Syndrome in various sheep breeds and some types of Dwarfism in various cattle breeds.
>
> **Complex inheritance:**
> **everything else not covered by simple inheritance.**
> Examples include: bovine mastitis, bovine TB

This manual only concerns diseases with simple inheritance, using *homozygosity mapping* to uncover the genetic basis for disease.

With simple inheritance, one can think of the gene pool of the population having a single defective form of the gene, called a defective **allele**, alongside other functionally normal alleles. Two special cases of simple inheritance can be defined, **dominant** and **recessive**:

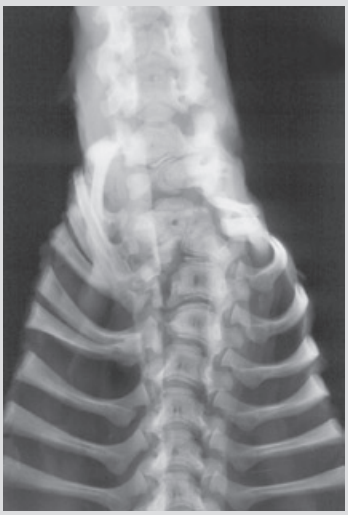
> **Dominant inheritance:**
> where inheriting one copy of the defective allele is always sufficient for the disease to occur.
>
> **Recessive inheritance:**
> if inheriting one copy of a **normal** allele is always sufficient to ensure normal form and function. Disease only occurs if an animal inherits two copies of the defective allele.

When an individual has two copies of the same allele it is said to be **homozygous** (if the two alleles for a gene are different, it is said to be heterozygous). Homozygosity mapping is therefore concerned with diseases that are recessive since it relies on the existence of homozygosity at the disease **locus** in *all* animals that have the disease.

Note that an animal carrying two copies of the defective allele need not imply that it will show the disease symptoms; however having the disease must imply the animal carries two copies of the defective allele. Therefore homozygosity mapping may be effective even if there are environmental factors that influence incidence: for example, **homozygotes** for the defective allele controlling susceptibility to a pathogenic disease will not acquire the disease if they have not been exposed to a disease agent. Congenital abnormalities (i.e. present at birth) are commonly associated with recessive diseases.

Why are recessive diseases a problem?

It is worth considering why recessive diseases are a problem for animal breeders, since they are at the heart of a common story. A breeder finds that one animal has a congenital disease, and to avoid publicity the animal is quietly disposed of – it is an understandable reaction. However the animal's siblings (sibs) are then used for breeding and the breeding success of a sib may spread the defective allele and the disease begins to get noticed, and everyone knows someone who knows of a case. For example, if a popular male happens to be a carrier of the defective allele and his daughters are retained for breeding it can become quite common. In a recessive disease the defective allele is a lot more common than it appears from the prevalence of the disease – **if 1% of the animals have the disease, 18% are carriers!**



X-ray showing the malformed vertebrae in an animal with CVM

Success Story: Finding the Founder

Most recessive diseases result when a mutation first arises in one founder animal which is then favoured for breeding. If used extensively, and the frequency of carriers can increase rapidly, resulting in the disease occurring when two carriers are mated. This process has been nicely demonstrated for a disease that was observed in Holstein cattle.

Complex Vertebral Malformation (CVM) is a recessive inherited disease that is associated with a high level of foetal mortality with skeletal malformations, particularly the vertebrae (see x-ray image opposite). Genomic analysis has identified the causal mutation in a gene that plays an essential role in mechanisms controlling the formation of vertebrae.

The origin of the defect has now been traced back to a single sire that was born in the USA in 1963 (*Penstate Ivanhoe Star*). One of his sons *Carlin-M Ivanhoe Bell*, who received the defective allele from his father, had high genetic merit for milk production and consequently founded a breeding line that was extensively used globally resulting in around 30% of AI sires available in many countries being CVM carriers.

Following the development of a commercially available test, Holstein animals have been routinely tested for carrier status with results published, particularly if they are destined to be used as AI sires.

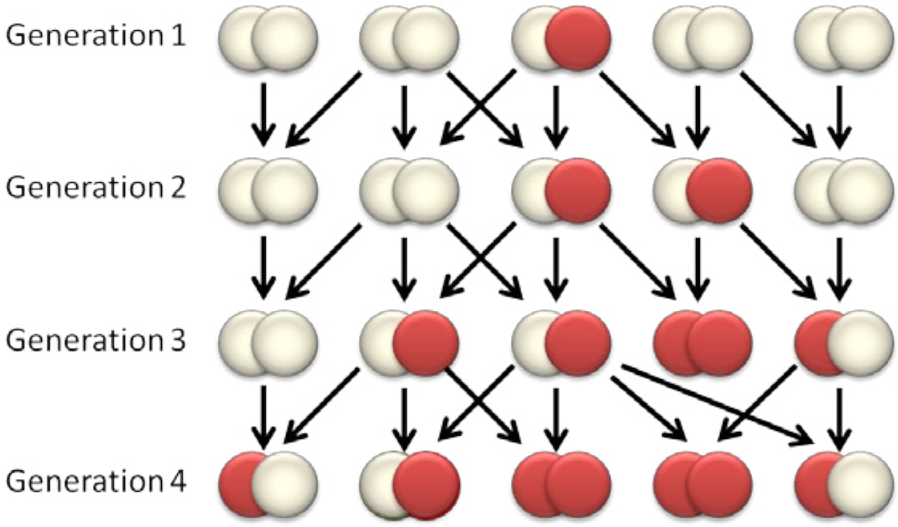
X-ray image courtesy of JS Agerholm, University of Copenhagen and Cold Spring Harbor Laboratory Press. Previously published in Genome Research 2006, 16: 97-105

The origins of the problem can be made clearer by examining **Figure 1**, where a small population is tracked over 4 generations.

Generation 1: A breeding individual appears carrying a single copy of a new defective allele (red), which is recessive. Since the individual still carries a normal copy of the allele (grey) it is fit and healthy. This individual has several offspring to which it can donate either a normal or a defective allele. The offspring have a 50% probability of inheriting the defective allele: in the figure below, 2 of the 3 offspring from the original carrier inherit the defective allele.

Figure 1: Tracking recessive alleles over four generations.

The red allele is defective and recessive. Normal alleles are grey.



Generation 2: Since there was only one copy of the red allele in generation 1 and animals have two different parents, each of the offspring inheriting a red allele (in generation 2) will only carry one copy and so will appear fit and healthy, and may be selected for breeding. There will be no disease observed in generation 2 because no animal carries two copies of the red allele.

Generation 3: The red allele may now increase further in frequency depending on the breeding success of the parents. It is possible that two of the carriers will be mated together, although in this early generation this is equivalent to mating two half- or full-sibs together, but it is feasible that the disease will occur. However it is very possible that it will not be recorded for two reasons: (i) the probability of an offspring of two carriers having the disease is only ¼ since the offspring must inherit the red allele from both parents and this happens with probability ½ x ½; and (ii) even if the disease occurs it may not be recognised, as it is a new disease, and even if it is recognised it may be attributed to other causes, or perhaps just ignored and disposed of. Meanwhile in generation 3, other carriers are breeding and spreading the red allele further in the population. Consequently the red allele may become yet more frequent, and the problem may still not have been recognised, although many individuals are carrying the defective allele.
Generation 4: The disease is increasing in prevalence and may begin to be noticed

and recorded. At this stage, there are many individuals carrying a copy of the defective allele, increasing the chances of more animals in further generations being affected by the disease.

NOTE: the frequency of carriers is far higher than cases of the disease; so for a disease prevalence of 1%, it would suggest a frequency of carriers in the breeding population of 18%.

Mutations that result in defective alleles are happening all the time, yet few of these become problems. Figure 1 shows that for the disease to be observed the same ancestor must occur on both sides of the pedigree, and, to this extent, recessive diseases are associated with the process of inbreeding. Consequently there are two breeding processes that promote the spreading of a defective recessive allele; high rates of inbreeding and positive selection:

Inbreeding: high rates can occur when either:

a. A small number of males or a small number of females are used as parents (note a large number of females does not make up for a small number of males in terms of controlling inbreeding).

b. Some individual parents, for example bulls, rams or stallions, are allowed to contribute large proportions of breeding offspring to the next generation. This may occur in less numerous breeds or even in numerous breeds where popular sires may dominate breeding (for example, through use of Artificial Insemination).

Positive Selection: can promote higher rates of inbreeding unless properly managed and there are two processes in particular that may act to further promote the occurrence of a recessive disease emerging:

a. One copy of the defective allele may help the individual perform better and hence give it a selective advantage. In this case carriers of the defective allele will be favoured in selection.

b. The gene with the defective allele may be situated very near to a gene which has a big impact on the variation in the desired trait, and the deleterious mutation may have occurred on the same bit of DNA carrying an allele that has a beneficial effect on the trait. Consequently selection for the beneficial trait allele increases the chance of spreading the defective allele through a phenomenon known as **linkage** (explained further in Step 2).

Preventing the diseased animals from breeding does not remove the problem. The key to removing a recessive disease is to also prevent the carriers from breeding. As carriers can often appear fit and healthy, they are difficult to identify, which poses a problem. However, as a result of recent advances, DNA technology may offer a fairly low cost route to identify carriers of recessive diseases and therefore remove them.

Step 1

Crooked tail - the consequence of a carrier advantage

Carrying a single copy of a defective gene can sometimes result in a breeding advantage or a more desirable phenotype. When this happens, it can result in a rapid increase in the number of carriers within a breed and therefore in the incidence of disease. Such a problem has been identified in Belgian Blue cattle where the carriers of Crooked Tail Syndrome (CTS) have been shown to exhibit greater muscling.

Crooked Tail Syndrome (CTS) is a condition that as well as causing a characteristic crooked tail, causes affected animals to display slowed growth from about 1 month of age onwards, have an abnormal skull type (shortened broad head) and extreme muscular hypotrophy. On occasion animals can also exhibit; short, straight and extended front legs, curvature of the spine and asymmetrical muscle development of the back. Although not lethal, most extreme cases tend to be euthanized on welfare grounds or result in economic loss for producers due to poor growth or carcass depreciation.

The incidence of CTS was noted to have increased very sharply in the Belgian Blue breed around 2008 and retrospective analysis revealed that around 25% of animals in the breed appeared to be carriers. It was later found that carrier animals tended to be more heavily muscled, have thinner skeletons and more rounded ribs – preferred characteristics for breeders, which may have explained the reason for the sudden increase in frequency.

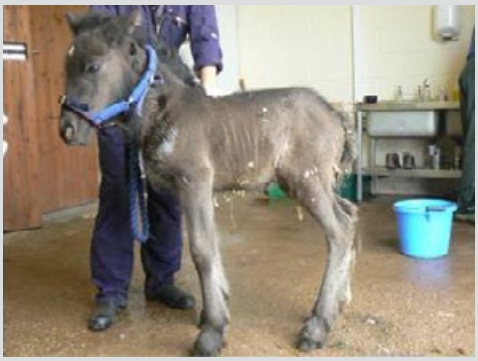
A genetic test for carrier status has now been developed which allows breeders to manage the disease in the Belgian Blue breed and to work towards elimination of the defective version of the gene.

Belgian Blue with Crooked Tail Syndrome exhibiting high muscling. Photo courtesy of Arnaud Sartelet and Carole Charlier, University of Liege – work undertaken as part of the "Rilouke" project, Towards comprehensive management of genetic defects in Belgian Blue cattle breed (BBCB), funded by the Walloon Region". Source: Fasquelle et al. (2009) PloS Genetics Vol 5, issue 9

Step 2:

Understanding homozygosity mapping

The difficulty with recessive diseases lies in not being able to identify carriers; if these were readily identifiable then it would be straightforward to remove the defective allele effectively. If one knew which of the many genes in the genome had the defective allele, and what the mutation was that made the allele defective, then a simple DNA test could be designed and animals could be screened for the defective allele.



Fell Pony foal with FIS.

Success Story: Homozygosity mapping in action

The homozygosity mapping approach has already been successfully used to develop a genetic test that breeders can use to identify carriers for a devastating disease in horses.

Foal immunodeficiency Syndrome, formally known as Fell Pony syndrome, was first recognised in UK Fell pony foals in 1998 and was subsequently reported across Europe and the USA. In 2008 the same condition was also confirmed in the Dale Pony breed.

The disease manifests at 2-6 weeks after birth, foals being apparently normal when born. The first clinical signs of the disease include diarrhoea, nasal discharge, poor growth, pale gums and decreased appetite. Foals also develop a progressive severe anaemia. The syndrome has a 100% mortality rate.

Analysis of the pedigree revealed that the Fell Pony breed had undergone a genetic bottle neck in recent years, resulting in a small effective breeding population, which suggested that the disease could be a recessive inherited disease.

By following a homozygosity mapping approach, researchers at the Animal Health Trust and the University of Liverpool were able to identify the causal mutation. The work allowed an effective genomic test to be developed in November 2009 that can be used in both the Fell Pony and Dale Pony breed. Results from subsequent random screening have suggested that 38% and 18% of the UK Fell Pony and Dale Pony populations respectively carried the defective gene. The availability of the test will allow breeders to effectively manage the disease and work towards its elimination.

Photo courtesy of Dr. Sarah Blott, Animal Health Trust.

Step 2

Discovering the causative mutation has until recently been an expensive and time consuming process, and one that requires significant numbers of cases. This has made it beyond the reach of all but the richest breed societies and breeding companies, with the defective allele already well established at significant frequencies in a population. Recent advances in DNA technology may offer a new, affordable opportunity to address the problem using *homozygosity mapping*.

Homozygosity mapping: is based on the idea that for any recessive disease, the one thing that all cases have in common is that they all have two copies of the defective allele. With a large enough number of cases it would therefore be highly significant to see complete homozygosity for the same allele in cases, yet variation in the **controls**. If we were to sequence the genome of a number of cases and a number of controls then this homozygosity will be observed in the cases but not in the controls. One can go further if the disease is entirely under genetic control (i.e. no environmental factors): in this case *none* of the controls would be expected to share the homozygosity seen in the cases – although they may be homozygous for a different allele. Surprisingly even though there are 3 billion **nucleotide** bases in the genome, with as few as 5 cases and 5 controls, it is unlikely this would occur by chance. This approach therefore makes it possible to identify the position of the mutation with relatively little effort through the use of SNPs.

The approach taken above assumes that it is possible to sequence the animals, which although feasible is still costly. However, the same result can be achieved using **dense SNP chips** (i.e. chips with a high number of markers (SNPs) spaced evenly along the genome), taking advantage of the phenomenon of linkage where bits of DNA close to each other on the same **chromosome** (the structures that the DNA in our genomes is organised into) tend to be inherited.

Sharing the whole chromosome is unlikely due to a random process called **recombination** which will stop this occurring; however a large segment around the defective allele is likely to be shared by the carriers for a number of generations. This will be repeated each generation, so when an individual inherits two copies of the red allele it will also inherit two identical copies of the DNA either side of the gene. Therefore there will be a segment of DNA that is completely homozygous in all the cases, not just the causative mutation. This process is illustrated in **Figure 2**. Over time, successive recombination events would be expected to reduce the size of the homozygous region, but in most cases this would be expected to be a relatively slow process.

The larger the segment of homozygous DNA, the larger the target and the more likely it can be found using a SNP chip. Although the length of this segment will decrease over time, with a sufficiently dense SNP chip, this homozygosity will be observed even a number of generations after it first arises. This will *not* occur in controls. Current technology is delivering ever more dense SNP chips making it reasonable to expect to observe a segment of homozygosity for most recessive diseases in domestic animal species. It is this idea that was tested by the RIDGENE project and is described in this Manual.

Example:

Species	SNP Chip Density Available*	Species	SNP Chip Density Available*
Cattle	Up to 700k	Pigs	Up to 60k
Chicken	Up to 700k	Equine	Up to 60k
Sheep	Up to 700k	Dogs	Up to 170k
Goats	Up to 60k	Cats	Up to 60k

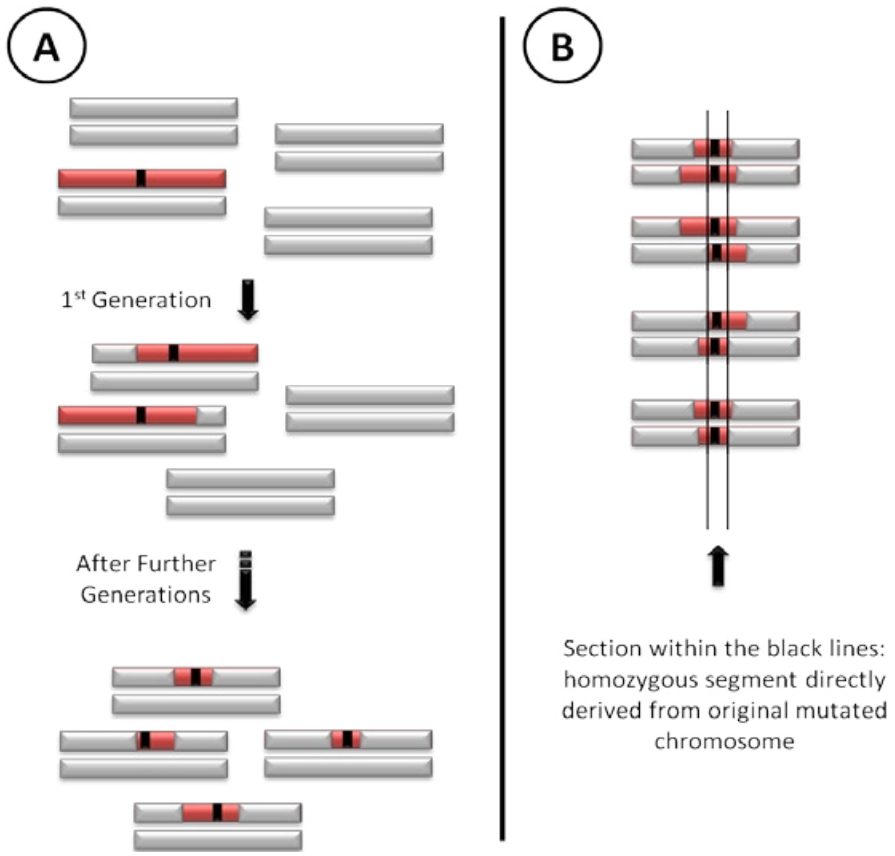
*correct at time of publication

It is this idea that was tested by the RIDGENE project and is described in this Manual.

Figure 2: An illustration of how homozygous segments arise

A: a deleterious mutation (black) occurs in the red chromosome, with **grey** chromosomes representing all other normal chromosomes. In the first and later generations, recombination occurs to break down the intact segment containing the mutation. The area of **red** chromosome surrounding the mutation gets smaller over successive generations.

B: the carriers are mated together producing four cases which each have a small homozygous segment containing the mutation.



Identifying the existence of a recessive disease and the need to develop a test

There is no simple way of identifying a recessive disease but there are indications and contra-indications. In general it is possible to be more confident for species that have large litters e.g. pigs, and harder for species or breeds that typically give birth to singletons such as cattle and equines.

Figure 3 provides a simplified decision tree for looking at this problem. The notes on the questions asked in Figure 3 (shown in the supplementary box) are intended to give guidance on the reasoning behind them.

NOTE: Carefully following the decision tree to determine whether the disease is likely to be a recessively inherited condition before taking further action will save both valuable time and money and will help ensure that the most appropriate course of action is pursued.

If it is concluded that a recessive disease is a possibility, it is worth bearing in mind that there are many complex forms of inheritance and until the genetics are understood there can be no certainty in this conclusion.

An additional test for species with large litters:

For species with litters rather than singletons it is possible to take a further step based on the observation that even if 2 carriers are mated together the probability of an offspring being affected is only ¼. Therefore in a litter with 12 offspring, 2 carriers mated together might be expected to produce 3 affected animals in a single litter, sometimes 1 or 2 or 4, but rarely 9.

The distribution of numbers affected in a litter can be tested for consistency with recessive inheritance. Reasonable confidence for this form of inheritance can be obtained with a relatively small number of large litters, whereas with litters of twins, large numbers of litters are needed to provide such confidence. This is described in detail in Step 7 (page 23) of this manual.

Figure 3: Decision Tree

An indicative decision tree for deciding whether or not a disease is caused by a recessive defective allele. The superscript numbers refer to the notes in the box on the following page.

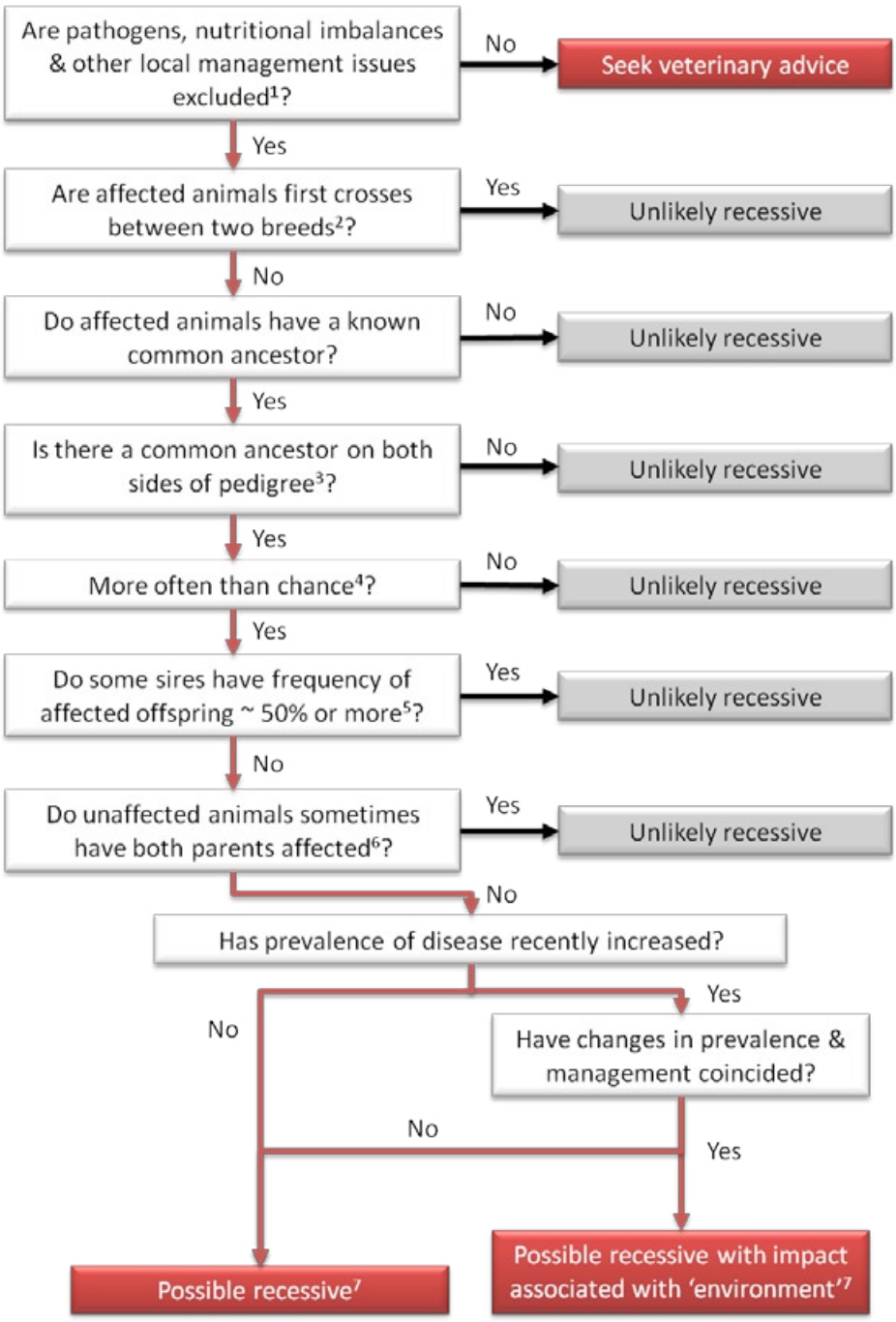


Figure 3: Explanatory notes

1. It is quickest and cheapest for the veterinary surgeon to rule out pathogenic agents associated with the symptoms, check the nutrient balance, seek for any contaminating toxins, and review the local management before assuming the disease is genetic.
2. Homozygosity relies on a common ancestor being present on both sides of the pedigree, and if the cases are first crosses between two breeds the occurrence of a common ancestor is likely to be a long time ago, and the possibility of a recessive disease becomes very unlikely. If the two breeds themselves are closely related (e.g. Beltex and Texel sheep) then the chances are increased.
3. Some ancestors being more common than others in the pedigree of diseased animals is evidence for a genetic association. However for a recessive disease, the same ancestor needs to be identified on both sides of the pedigree for all cases, and the same ancestor needs to be common to all cases.
4. The identification of a common ancestor on both sides of the pedigree of all cases becomes more likely by chance the further back in the pedigree that the ancestry is traced. This is a natural consequence of maintaining a breed. Therefore the identification of a common ancestor in (3) above must be assessed by examining how likely randomly chosen individuals from the breed share a common ancestor on both sides of the pedigree. This can be assessed by taking repeated random samples of the breed, with sample size equal to the number of cases, and seeing how often by chance that such a common ancestor can be observed either the same number or fewer generations back as was observed in the cases. This probability can act as a significance test of the finding.
5. When there is a recessive disease, with prevalence p, the defective allele has frequency $\sqrt{p}$ in the breeding population, and a carrier sire will only have a fraction $\frac{1}{2}\sqrt{p}$ of its offspring affected, assuming mating is at random. This is because the sire must pass the defective allele, which happens with probability ½, and the offspring must receive a defective allele from the dam, which happens with probability $\sqrt{p}$. Therefore it is very unlikely that ~50% or more of a carrier sire's offspring will be affected (since p will need to be 1!). Even if the prevalence is as high as 16%, only 20% of the offspring will be affected. Therefore some genetic process other than a recessive defect is involved.
6. For a simple recessive disease, if both parents are affected then both parents must be homozygous for the defective allele. If so, the offspring must inherit two copies of the defective allele and must be affected. Therefore if it is possible for an unaffected offspring to have both parents affected, the disease cannot be controlled by a single recessive gene.
7. If none of the contra-indications listed above apply then it is possible that the disease is controlled by a single recessive gene and the application of homozygosity mapping may provide a means to solve the problem efficiently.

NOTE: Successful homozygosity mapping relies upon true cases being identified, with no false positives.

Homozygosity mapping and epidemiology

Epidemiologists have two parameters that are used to assess the quality of diagnosis of a disease; sensitivity and specificity:

Sensitivity:
is the probability that a truly diseased animal is diagnosed as having the disease.

Specificity:
is the probability that a truly healthy animal is diagnosed as being healthy.

A perfect diagnostic test has both sensitivity and specificity equal to 1.

In homozygosity mapping, a healthy animal that is diagnosed as being diseased will almost certainly hinder the finding of a homozygous segment, as it does not carry two copies of the defective allele. Therefore the approach only works if the diagnostic test for identifying disease has perfect specificity i.e. no healthy animal is identified as having the disease. Homozygosity mapping will still work if the diagnostic test has less than perfect sensitivity, since it is the shared properties of the cases that are critical, and missed cases or cases treated as controls will not disrupt these properties. This is illustrated in **Figure 4**.

In conclusion, when sampling, always err on the side of excluding any disease cases that may be open to doubt.

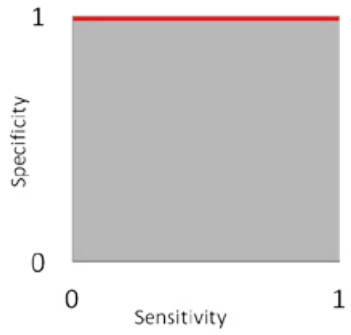


Figure 4: An epidemiological perspective on when homozygosity mapping is feasible.

The feasible options are shown in red and infeasible in grey.

NOTE: Successful homozygosity mapping relies upon true cases being identified with no false positives. It is strongly emphasised to either sample only cases that are well documented as being cases, or note on a summary sheet whether a case is well documented or open to doubt.

Doing the field work - Collecting DNA samples

What samples are needed from the field?

If it is decided that a recessive defect may be a possible cause of the disease being observed, then the most important task is to make sure that DNA is sampled from as many cases as possible. However paying attention to a number of key points when sampling will help make the outcomes more certain and more far-reaching.

The following animals should be sampled for DNA:

(i) All cases.

(ii) All dams of cases. Dams are a *very significant aid* in homozygosity mapping as the dam must have one defective and one normal allele. If the dam is unavailable for sampling, then sample an unaffected full-sib, or an unaffected half-sib, or an unaffected herd-mate.

(iii) All sires of cases.

(iv) A sample of 20 healthy animals from herds or flocks with no records of the disease.

Although predictive **genotypes** for carriers can be obtained with very small numbers of cases and controls (~8 each), as many cases as possible should be sampled to increase the chances of finding the causative mutation; ideally samples should be collected from cases as they occur. Having samples from the sires and dams of cases available is useful because they can help define a set of markers that can be used to identify carriers as well as potentially narrowing down the region of homozygosity to help locate the causative mutation.

It is sufficient to sample only the cases at the very early stages and to gather the **controls** at a later date, although this minimal strategy runs the risk of particularly useful samples from dams and sires becoming unavailable and lengthening the process. In some populations a number of animals may be genotyped with dense SNP chips already, which could potentially be used as unrelated controls – thus reducing the amount of sampling needing to be done.

NOTE: It is recommended to proactively **biobank** DNA samples from cases as well as from sires and dams (controls), when an inherited disease is suspected as it can help accelerate the process if homozygosity mapping is pursued. Given the recent advances in DNA technology and bioinformatics, the feasibility of obtaining such samples should be considered as part of the veterinary routine.

NOTE: Developing a reference set of animals, which have been genotyped with dense SNP chips, can act as general controls for a breed. This is beneficial even if their identities are not known. It is recommended to check whether such a reference set exists in your species/breed of interest as this may save time and money.

What samples should be taken?

The key requirement is that DNA can be extracted of sufficient quality and in sufficient quantities. Whilst quality and quantity may be less critical for genotyping, DNA sequencing, which plays a role in finding causative mutations in homozygosity mapping, requires high quality DNA. We would recommend that DNA samples that are of high enough quality to allow sequencing as well as genotyping should be obtained in order to find the causative mutation if desired.

As it's often very difficult to go back and take a second sample from cases, it is better to take appropriate samples at the outset. To help ensure sampling is effective the following table contains a list of products that have been shown to be effective for the purpose and others which are not: note that those that are ineffective for homozygosity mapping may be ideal for other purposes.

NOTE: The quality and quantity of DNA from hair (unless it contains the follicle) and from FTA cards is insufficient, and we do not recommend their use for this purpose. Not all nasal and saliva swabs are suitable as they are designed to diagnose infections by promoting the growth of bacteria and in the analysis the animal DNA is swamped by the bacterial DNA.

There are a number of viable easy-to-use alternatives such as:

Species/Type	Sample (including example where appropriate)
Cattle	Nasal swab kits e.g. Performagene kits (DNA Genotek Inc.)
Horses	Nasal swab kits e.g. Performagene kits (DNA Genotek Inc.)
Sheep	Saliva swab kits e.g. Oragene (DNA Genotek Inc.)
Goats	Saliva swab kits e.g. Oragene (DNA Genotek Inc.)
Sires in AI	Three semen straws
Poultry	1ml of whole blood in EDTA
Mammals	5ml of whole blood in EDTA
Dead Animals	Tissue such as ears

All samples should be labelled with:

i. An ear tag number and/or a herd book number
ii. Whether it was a case or control including relationship to the diseased animal
iii. A date and herd identifier (and/or post code)

All these details should also be written down and sent by post or by email to the project co-ordinator for the industrial partner (e.g. breed society or breeding company) and/or the contact point for the service provider (see Step 5).

Who can take samples from animals?

Given a satisfactory consideration of whether the disease in question is a recessive condition or not, homozygosity mapping falls under the Veterinary Surgeons Act due to its effectiveness in characterising the segment of DNA harbouring the defective allele, and the consequent identification of carrier animals. Therefore a vet is permitted to take blood samples from live animals for the purpose of a diagnostic test since the test will yield results that will be fed back to the breeder that will be directly beneficial for husbandry purposes. Nasal and saliva swabs can be taken by farmers for this purpose and may also have the benefit of costing less although buying the sampling kit may incur a cost. Tissue samples can be taken from dead animals by farmers or vets.

How are samples stored?

Kits will have their own instructions, however a guide to sample storage is provided in the following table.

Sample Type	Storage Requirements	Storage Duration
Semen	May be stored in a normal freezer for the purposes of DNA sampling (note however that the straws will no longer be fit for insemination)	Can be stored for months
Tissue	Normal freezer	Can be stored for months
Blood	Refrigerator	Should be transported promptly
Nasal (e.g. Performagene kit)	Room temperature	Can be stored for months
Saliva (e.g. Oragene kit)	Room temperature	Can be stored for months

When and how should samples be transported to the service provider?

It is assumed that samples will be sent to a service provider for analysis. It is recommended that samples are not stored on farms for long periods without transfer to the service provider.

Transport will normally be done by post. The postage label should clearly identify a designated contact person within the academic partner (e.g. f.a.o. A. N. Other). A summary sheet should accompany the samples detailing the information on the samples to act as a checklist for quality control. Please note that samples sent from outside the UK will require an import licence from Defra.

In order to preserve high quality DNA in samples from semen, tissue or blood, they should be packed in dry ice before sending through the post to keep them cool or frozen. Service providers can help arrange this. A big advantage of the new nasal

and saliva swab kits is that the quality of the DNA can be maintained in transit without special precautions since they are best kept at room temperature and should not be chilled or frozen.

As well as the samples, additional information that should be collected and provided to the service provider includes:

- Contact details for the vet who attended the case and any post mortem results (if obtained).
- A photo of the diseased animal (as in some cases a photo can help to define and characterise the condition) – vets may be best qualified to take photos that show distinctive features of the condition.
- The results of a post mortem examination of a small number of cases (e.g. 3) if possible (as the reports on the primary lesions could assist in identifying the causal mutation).

How many cases need to be sampled?

In the pioneering work on homozygosity mapping, a causative mutation was found with only 3 cases. However these guidelines recommend the accumulation of 5 cases and associated controls before starting the laboratory work.

We would recommend continuing with the collection of samples until 20 cases and associated controls have been sampled, and then review with the service provider whether homozygosity mapping is the right approach and whether sampling strategies need to change.

Working with a service provider

In most cases carrying out homozygosity mapping will require partnership with a service provider to genotype the samples and process the results. The industrial partners should expect the following from the service provider:

- An indication at an early stage on whether or not the condition is a recessive disease, with a putative chromosomal location. This may be feasible after 5 cases and controls.
- A small number of SNP markers that can be genotyped routinely, with instructions on how to identify carriers from the genotypes. Routine genotyping for a small number of SNPs is affordable as a screening programme and this has been implemented in one of the case studies carried out.
- Details of the causative mutation (if requested) allowing carriers to be identified, and a test to identify carriers rapidly and cheaply. This may require more than 8 samples to do cost effectively and will require DNA that is of sufficiently high quality, although the quantity required is not large.

If the disease is identified as having a recessive inheritance then actions can be taken as described below. If it is not a recessive disease, then other approaches will be necessary and discussions with service providers and the Biosciences KTN are recommended to develop a plan for how best to proceed given the epidemiological, pathological and genetic information gathered to date.

How to engage with a service provider

You may already know a suitable service provider you could work with. However, if you do not, the Animal Sector team within the Biosciences KTN would be able to help you with this, and can be contacted as follows:

Email: **info@biosciencektn.com**

Telephone: **+44 (0)131 651 7334**

Please mention that you would like to discuss the “RIDGENE manual” or “Inherited diseases of animals”.

Testing for carriers & managing disease

In principle, once a test for a carrier has been identified it is straightforward to remove the disease; however, it is advisable for this process to be carefully managed. Managing a breed's gene pool is the collective responsibility of all its breeders. This means that at the outset of any scheme, before results are known and targets are set, there should be an agreement between breeders for ensuring openness and ready availability of genotyping results.

Lack of openness in the occurrence of diseases is a major contributor to the spread of recessive disease. Homozygosity mapping provides a cost effective opportunity for nipping such diseases in the bud, but it depends on the readiness of breeders to alert the breed society to the possibility of a disease and the support of the breed society in responding positively.

The risk of removing a recessive defect from a small population, by rapid selection against carriers, is that favoured blood lines can be severely and unnecessarily curtailed and an overemphasis on preferred normal sires may actually increase the frequency of other deleterious alleles that are likely to be present in the population.

Therefore, in developing any scheme to remove inherited disease, the following points should be carefully considered:

1. **New cases:** It is necessary to be alert for new cases and to check that they conform to the expectations until the causative mutation is identified with confidence

2. **Build confidence:** You can gain confidence in the outcome of the analysis by:
   a. using the control samples collected from herds or flocks with no record of disease (see *Step 4: What samples are needed from the field*, point iv.) to demonstrate the (near) absence of the carriers
   b. showing the sires of the cases are carriers

3. **Check the frequency of the defective allele:** The most effective next steps will depend on how widespread the defective allele is in the population. Since breeding males are often a highly selected group of animals the estimate of allele frequency is best achieved by genotyping a random sample of 50 breeding females with ages stratified to represent the female breeding structure. Random sampling should ensure that blood lines are represented according to their contribution.

   With 50 breeding females genotyped, the **standard error** (s.e.) of the frequency estimate should be reliable: ~ 0.04 if the allele frequency is as high as 0.2. The carrier frequency in the population is also estimated with an approximate s.e. of 0.07 if the carrier frequency is as high as 0.5.

4. **Check the link to selection objectives:** It is useful to see if the selection objectives are associated with carrier status, as the defective allele may be associated with desired performance, and may be closely linked to a favoured quantitative trait locus (QTL) that is perhaps unrecognised. If males are used in AI, genotype these males and compare the carrier frequency with that

found in the females, and test for differences in performance (or EBV) between carrier males and normal males. However, please note that in small breeds it is possible that there will be insufficient power to obtain a definitive result.

Removing the disease

An effective way of removing the disease is to allow only non-carrier males to breed, or be registered in the herd book. In this way no offspring can inherit 2 defective alleles. Over time the frequency of the defective allele will diminish to a level that is sufficient to allow non-registration of females. The benefit of this slower approach to removing the allele is that it gives an opportunity for favoured characteristics of blood lines that have high carrier frequencies to remain in the population (unless of course the defect is very strongly associated with such characteristics).

In some cases the carrier frequency may be very high and the number of normal sires available is too small for rapid removal whilst maintaining sustainable breeding policies. One test of this is to examine the number of non-carrier sires entering the breeding population per generation in the period before the test was available. If this number was fewer than 25, then targets need to be set for a phased reduction in the allele frequency over time, with breeders advised beforehand of the reductions. This may be achieved by limiting the number of carrier males being registered, perhaps by quotas applied to the breeders. If such a policy is necessary then it is also necessary to ensure that carrier sires are mated to normal females. Not only does this avoid disease but it also increases the chances of their offspring not being carriers. This approach requires genotyping the females.

Step 7:

Testing recessive inheritance in the lab

The lab analysis is initially concerned with isolating a set of markers for identifying the carriers as cheaply and as rapidly as possible, before proceeding towards the discovery of the causative mutation (if required). To achieve this, it is assumed that the first step is to carry out a high density genome scan using a SNP chip to identify the homozygous segment carrying the defective allele. The cost of the SNP chip is assumed to be cheaper than sequencing, which will be true for the medium term. This approach makes no *a-priori* assumptions on candidate genes that may harbour the defect. Whilst targeting candidate genes may appear to offer a short cut method to detection, there are many occasions where assumptions concerning candidate genes have been proved to be wrong, thus resulting in wasted effort and resources. Using a high density scan is comparable in cost and speed to testing a relatively small number of candidate genes.

It is not necessary to write a lab manual on how to do the lab procedures, since these are standard services; therefore this section takes the approach of giving a description of the sequence of processes to be undertaken by the service provider, including a more technical appraisal of some previously mentioned aspects.

Testing recessive inheritance in species with large litters

When the species has large litters as in pigs, it is possible to test for departures from recessive inheritance, which is a simple and useful first step (as described in Step 3: 'Additional test for species with litters'). When two carriers are mated together then the probability that an offspring is affected is ¼. Therefore providing the defect does not affect development to term, and littermates are independent of each other (which will be the case if the disease depends only on the inherited genome) the distribution of the number of diseased offspring in a litter of size n is binomial B(n,¼). Therefore with a number of affected litters it is possible to build up a likelihood test. There is one complication in that litters produced by two carriers have a probability of (¾)n of producing no affected animals and will not be detected. Therefore what is observed is a truncated binomial distribution with no zeros B0(n,¼).

The probability of observing j cases in an affected litter of size n is:

If $p = ¼$, the likelihood of observing j1…jm cases in m litters of sizes n1 … nm is proportional to:

From this a log likelihood test can be developed for H0: $p = ¼$ against an alternative H1: $p \neq ¼$.

In the pig case study undertaken, it was possible to show in a convincing fashion that the disease behaved as if it were a recessive disease, however the litter sizes were large (~10) and the number of litters was also quite large. For sheep and goats it is possible to have a flow of twins from affected families, but the power is much weaker: the truncated binomial B0(2,¼) is equivalent to observing 1 case and 2 cases in twins with probabilities 6/7 and 1/7 respectively and in excess of 100 affected litters are required to obtain a 95% confidence interval of approximate length 0.2 spanning ¼.

Assessing the significance of homozygosity findings

Early work on homozygosity mapping used allele frequencies at the loci within the homozygous segment to assess the significance of a finding, making some assumptions on the independence of these sites. There are two sources of problems with this approach, first the estimation of frequencies from selected samples, and secondly the incorrect assumption of mutual independence for neighbouring loci. This manual recommends an alternative approach described below, although technology will no doubt advance to remove the problem (see 'Flowing from homozygosity to mutation' below).

For any individual taken at random from the population, there will be runs of homozygosity in the genome with a maximum length. For any two individuals taken at random only a fraction of these runs will be shared and therefore the maximum shared length of homozygosity will decrease, since by chance individual genotypes within a run will differ. In this way, as the number of randomly sampled individuals increases so the maximum length shared will decrease. However, if those individuals have a recessive disease in common, all the individuals will share the homozygous defective allele, and whilst the length of the homozygous shared segment will decrease, it will do so much more slowly than in other parts of the genome. This leaves open the question, how long might one expect a shared segment to be at random?

This question is difficult since the length of shared segments will depend on the set of relationships between the cases; more related animals may be expected to share more homozygosity because of their shared breeding history. The first step to overcome this is to compare the maximum shared run of homozygosity found on different chromosomes in the set of cases, since the breeding history will be identical chromosome by chromosome, if judged by pedigree alone. On the chromosome carrying the defective allele, the shared homozygosity will arise from more than this shared breeding history. Therefore with a set of 29 chromosomes, there are 29 measures of shared homozygosity and the question becomes one of detecting an apparent outlier to this distribution.
In this way it is feasible to find candidate segments through the set of cases alone. Whilst this is a step forward there are problems and the most important of these is that shared homozygous segments among cases can arise through selection on QTL other than the defective allele, leaving what is called 'a selection footprint'.

> Example: in one breed of dog it has been observed that a sample of 10 healthy individuals might very easily share 20cM of homozygosity on a particular chromosome due to its role in selection for a particular conformation. Therefore in the absence of controls observing such a shared segment would be highly misleading.

There are further issues:

- Chromosomes will be of different lengths and this will introduce some differences in expected shared homozygosity
- The properties of the markers collected on a dense SNP chip will introduce variation: if markers are less dense on a chromosome A than on B it may be more likely for A to show a greater length of shared homozygosity as measured by cM, but on the other hand it is more likely for B to show a greater run of shared homozygosity as measured by markers

- The markers chosen in one region may have a different distribution of frequencies, or a different **linkage disequilibrium** (LD) structure to markers chosen in another. If a segment has markers with low minor allele frequencies it is *a priori* more likely to appear homozygous by chance. The issues associated with the choice of SNP markers appearing on chips will diminish as chips become denser and more refined in marker selection.

For these reasons it is advisable to compare the cases with a set of controls taken from the same gene pool, and with as much shared breeding history as possible. This eliminates, to a large degree, the impact of selection footprints and characteristics of the marker sets used to assess homozygosity lengths. The shared breeding history of the cases argues that the cases may still have longer runs of shared homozygosity than controls, and identifying the segment harbouring the shared, defective allele require s the seeking of an outlier from the relationship between homozygosity runs in cases and runs in controls segment by segment.

Sourcing controls

Inferring the significance of the homozygous segments found in cases will depend on the source of controls.

(i) **Dams**: This manual recommends selection of dams of cases as the primary sources of controls. Dams provide the strongest evidence of interaction between the homozygosity of a segment, and whether an animal is a case or a control. This is because dams are expected to display no extended homozygosity in the disease segment, exactly where the cases are expected to display extended homozygosity. The dams share very similar breeding history although technically they are expected to be slightly less inter-related as a group than the cases. Therefore length of random runs of homozygosity may be marginally less than in the cases, but this is unlikely to be large or important.

(ii) **Random samples**: An alternative to using dams as controls is a random sample of individuals from the breed genotyped using a dense SNP chip for other projects. This approach was used in one of the sheep case studies. The strength of this option is that the costs of DNA extraction and genotyping will already have been met. The weakness is that the controls will tend to have less shared breeding history with the cases and more care is required in interpretation of outcomes. Sampling healthy herd- or flock-mates of the same breed is a further option. Here, the cases are likely to have more shared breeding history, but will incur DNA extraction and genotyping costs.

(iii) **Full-sibs**: In the pig case study, non-affected full-sibs were sampled as controls. This has a simple benefit in that cases and controls share the full breeding and selection history, differing only in the inheritance of two copies of the defective allele. Consequently statistical inference can be made exact by the use of permutation testing i.e. testing significance by randomly permuting the case and controls within full-sib pairs.

Two full-sibs have a probability of ¼ of sharing the same genotype by chance, and so comparing full-sib cases and controls will reduce the chances of declaring spurious shared segments among the cases as significant, as the same segments are also more likely to be shared with the controls. This benefit is small, and

vanishes as the numbers of cases and controls increase; therefore, this does not offset the benefit from using dams as controls, since dams must be heterozygous for the defective allele. Furthermore this option is not available for all species (i.e. those with small litter sizes or single progeny), and it is considered that sampling dams provides more informative controls.

(iv) **Sires**: are useful in the same way as dams and should be sampled as they always provide an informative contrast. Any test for carriers should always be tested on the sires, since sires must be carriers.

Practical tips in assessing significance

There are a number of useful checks that can be applied to check findings.

(i) Observing homozygous segments depends critically upon markers on dense SNP chips being assigned to the correct chromosomal location. An assignment error for a SNP location may split an unusually long homozygous segment into two that are typical. Such errors will diminish quickly for many animal species as the sequences become more refined through technology and through scrutiny. However, for the medium term, errors in the SNP location are an inherent problem. For this reason, whilst homozygosity runs should first be judged as observed, a second analysis should be carried out where a single marker breaking a run that is otherwise homozygous is ignored.

(ii) Testing should not be restricted to the maximum homozygous segment on a chromosome. As described above, it is possible for long homozygous segments to arise as selection footprints from other causes. As stated above, the defective allele is to be found in the discrepancy in the overall relationship of lengths of homozygous in the cases and controls when observed segment by segment.

(iii) A homozygous segment may harbour the recessive allele even though one (perhaps more) controls may share the same homozygous segment. There are at least three reasons why this may occur:

- The aetiology of the disease involves an environmental stressor and the control is a truly susceptible animal but was not exposed to the stressor;
- The animal was a true case but was misdiagnosed as being healthy
- At the time of the original mutation, there were segments that were identical to the one within which the mutation occurred, and these were carried by other contemporaneous ancestors. Descendants homozygous for the un-mutated segment may appear identical as judged by the markers on the dense SNP chip that is being used – unless the mutation is on the SNP chip which is highly unlikely.

(iv) It is feasible to carry out a Genome Wide Association Study (GWAS) as well as a test for homozygosity, to look for different marker allele frequencies in cases and controls. Such analyses on the small number of cases and controls, and with a recessive inheritance will lack power. Finding a positive result associated with a long run of homozygosity in the cases is an interesting observation, but should not be judged as being particularly meaningful.

(v) In all the case studies homozygosity was assessed primarily by marker runs not by distance. This was a subjective decision by the research team.

Defining a set of markers for selection against carriers

Once a homozygous segment is identified, a set of markers need to be selected for routine genotyping to define the carriers. The first step is to define the set of **haplotypes** found in the population that span the homozygous region using the cases and the controls. Apart from sex chromosomes, each animal carries two versions of each chromosome and a haplotype is the sequence of DNA on a single version, whereas the genotypes observed are the combination of these two versions (see the 'Moving from genotypes to haplotypes' box below). It is not straightforward to infer haplotypes from a long sequence of genotypes, but software such as FastPhase will assist in this process and derive a set of haplotypes. In the pig case study, the genotypes at the 62 loci of the homozygous segment were explained by 7 haplotypes. One of these haplotypes, the disease haplotype, was carried in homozygous form by all the cases.

Using the output from FastPhase it is possible to find a set of SNPs that differentiate carriers from non-carriers. In the pig case study, the homozygous segment observed in the cases extended across a region of approximately 8cM and over such lengths it was considered advisable to identify distinguishing loci at each end and in the centre of the segment so that it was feasible to discern recombinant animals. It was found that a total of 8 SNPs could be used for this purpose in blocks of 3, 3 and 2. The identification of SNPs discriminating carriers and non-carriers was done by observation and no algorithm was developed for the purpose.

After genotyping an animal for these 8 SNPs, it would be possible to identify it as being a carrier, a non-carrier or a recombinant. A recombinant animal is one where some crossing over between the two copies of a chromosome has occurred within an animal, and one block of SNPs may indicate a carrier but another block may indicate a non-carrier. The crossing over has resulted in a new haplotype being introduced into the population. This recombinant may or may not carry the disease, depending on where the crossing over has occurred. Using the recombinant may result in the disease, however it will also provide fresh information with which to reduce the length of the disease haplotype, and so help to identify the actual mutation. Therefore judicial use of a recombinant may be beneficial. The chance of recombinants occurring decreases as the length of the interval reduces.

Moving from genotypes to haplotypes.

For a SNP locus there are 3 possible genotypes: if the alleles are labelled '1' and '3' it is possible to have '11', '13' and '33'. For n loci then there are 3n possible sets of genotypes. In practice it is rare for so many sets of genotypes to be observed, since the alleles are inherited as single chromosomes with a sequence of alleles on each chromosome. With N parents there will only be at most 2N haplotypes for short segments of DNA. For example, with 3 loci, it may be that all the individuals in the population are described by only 6 out of the 27 possible sets of genotypes: (11,33,33), (11,33,13), (11,33,11), (13,13,13), (13,13,11) and (33,11,11). This is because there are only 3 haplotypes in the population '1 3 3', '1 3 1' and '3 1 1' and each of the observed sets of genotypes is a combination of two of these haplotypes.

Flowing from homozygosity to mutation

A decision tree to help select the appropriate methodology to use to move from homozygosity to mutation is shown in Figure 5. It requires a higher quality of DNA than the SNP genotyping, since all sequencing techniques rely on having longer fragments of DNA. The steps are expanded on below.

1. Sequences are not yet robust and error free. Therefore having identified a homozygous segment it is advisable to check the order of the known genes in this segment and 1cM either end of the gene. For example, different assemblies may have inverted the sequence order at the ends of the segment; if this were the case loci initially thought to be outside the homozygous segment may be seen to belong to the homozygous segment when inverted. If there is doubt over the sequence, extend the segment by 1cM either end.

2. Check databases for candidate genes although be aware that even short segments will contain many loci that might be considered as possible candidates. In the pig case study example, the 8cM homozygous segment identified contained at least 8 very reasonable candidates where documented function could be clearly linked to the observed phenotype.

Relevant databases include:

- Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim
- Online Mendelian Inheritance in Animals (OMIA), http://www.ncbi.nlm.nih.gov/omia
- Ensembl, http://www.ensembl.org
- National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov
- University of California, Santa Cruz, (USCS), http://genome.ucsc.edu

If, and only if, there is a candidate that gives a very, very close match to the observed disease phenotype then proceed to 'Sequence **exons**' in Figure 5 (see point 8 below). This makes use of a relatively cheap PCR technique that assumes the candidate region contains the mutation and the mutation is in the exon, however the cost benefits over other techniques that make no such assumptions are diminishing.

3. Many sequencing tasks require a reliable **reference sequence** for a species in order to reconstruct the DNA sequence for a new individual. If this is unavailable then it may be necessary to pursue the more time consuming and difficult technique called 'long range PCR' (see point 7 below). This is a continually diminishing problem as many species have a reference genome including: cattle, sheep, goats, horses, pigs, chickens, turkeys, ducks, and dogs. There is currently no reference genome for alpaca and salmon to name two species, but this will likely change over time.

4. There are a number of techniques that can be used for sequencing stretches of DNA. Their effectiveness will depend on length. With current technology the boundary between two important and powerful techniques 'array capture' and 'liquid capture' is around 5Mb.

5. For long segments it is feasible to use a technique called 'array capture'. Using the kit will sequence the entire segment for every animal. The cost of this technique increases in Mb blocks.

Example: in pigs, 1 kit will cater for 6.6Mb, 2 kits 13.3Mb. In practice, to maximise the chances of success the lengths sequenced should be restricted below this maximal length: in pigs 5Mb for one kit, 11Mb for 2 kits.

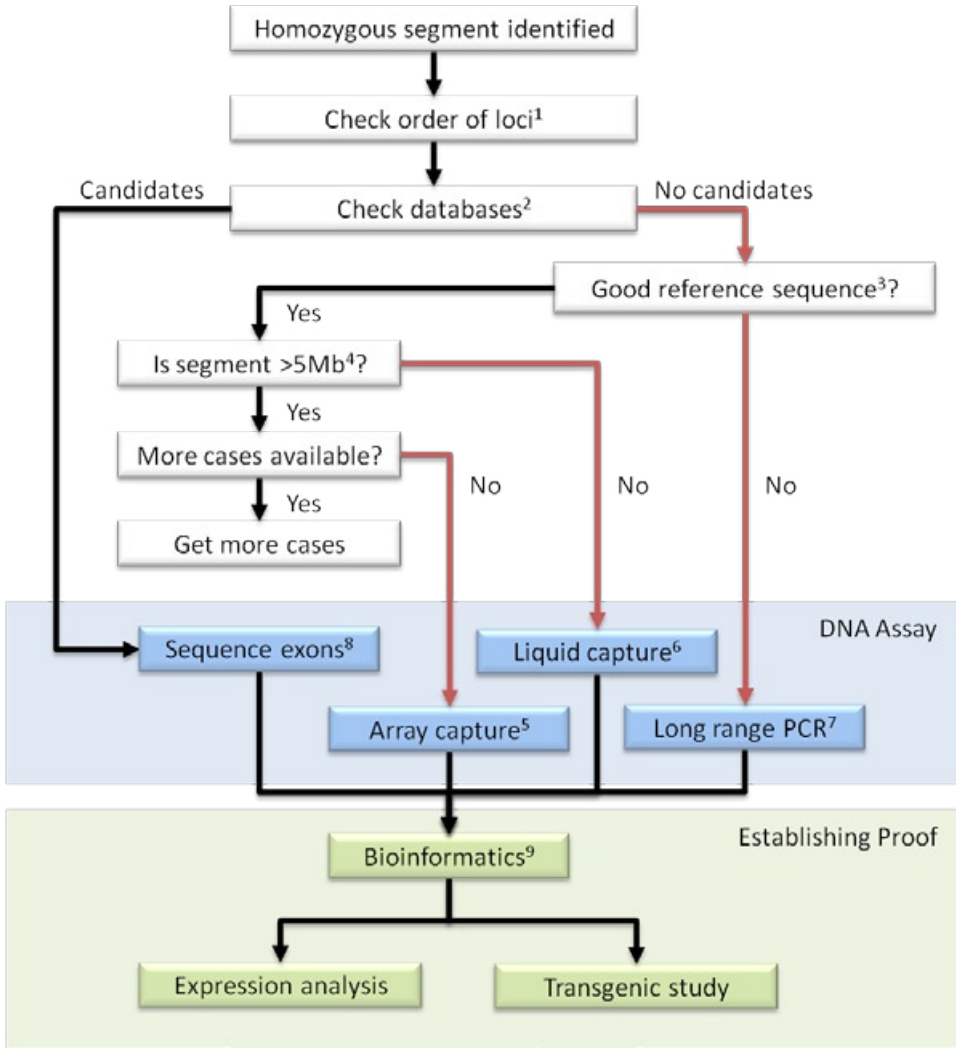
6. For shorter segments a technique called 'liquid capture' is preferable to 'array capture' since it is cheaper. As with 'array capture' the entire segment is sequenced for every animal.

7. Long range PCR is a long established technique for sequencing segments of DNA; however it is a long and arduous technique. This technique must be used if there is no reference sequence available.

8. If a particular candidate gene is strongly indicated and the gene has a moderate number of exons, then sequencing these exons using PCR is a viable, medium term option, being both relatively quick and cheap. The drawback is that assumptions have been made on the form of the causative genetic variant and other techniques are available that will sequence the whole segment.

9. The first step of the proof is that the variant identified coincides with the necessary conditions: primarily homozygous for all cases, heterozygous in parents, and rare and never in homozygous form in herds and flocks without the disease. The latter conditions hold unless there is an established environmental role in the disease which explains otherwise.

Figure 5: A decision tree for methodology to progress from a homozygous segment to the causal mutation. The superscripts refer to the notes provided above.



Step 8:

Best Practice

Technology Pitfalls

Although in principal homozygosity mapping is a relatively simple three step pathway: (a) observe a disease that appears to be recessive; (b) collect cases and controls to identify homozygous segment; (c) sequence to identify causal mutation, QED!, things can go wrong. This is a short section collating together some examples (many based on the experiences from the case studies) of what can go wrong, and steps that can be taken to minimise the risk.

1. The disease is not a simple recessive, and there will be no unusually long homozygous segments, or if such segments exist they will either be spurious, or will not lead to the removal of the disease. To protect against this, look for opportunities to test the assumption of recessive inheritance.

2. Sampling has been poorly controlled and documented so that healthy animals have been included as cases. To protect against this, ensure case definition is clear to all those collecting samples: encourage circulation of videos, photos, post-mortems, diagnostic indicators, contra-indications etc as appropriate.

3. Sampling procedures do not provide adequate DNA. This could be due to the biological sample obtained, storage prior to sending to those providing DNA services, or conditions of postage to the DNA service provider. To protect against this consult the earlier sections of the manual and check with the DNA service provider, and give clear instructions to those collecting samples.

4. DNA from cases and controls is adequate for use in SNP chips, but is not of sufficient quality for sequence technologies, necessitating further cases to be sampled. To protect against this consult the earlier sections of the manual and check with the DNA service provider, and give clear instructions to those collecting samples.

5. SNP order is wrong so that no unusually homozygous segment is identified from the SNP chips despite evidence of recessive nature and good sampling procedures. This will be a diminishing problem but may cause a problem in the medium term. An alternative explanation is that the causal mutation is very ancient and that the disease has been present in the population for a long time. To protect against this:

(i) Keep abreast of the latest sequence builds to ensure sequence order is most up to date
(ii) Use procedures that are not sensitive to the misplacement of a single SNP
(iii) Research breed history

6. A homozygous segment identified by SNP chips is too long for sequence capture technologies. This is becoming less of a problem as technology costs decrease. To protect against this collect as many cases as possible, and continue collection even though initial DNA analysis has begun.

Looking to the Future

DNA technology will get cheaper and more effective in homozygosity mapping. Technology is advancing rapidly, and has advanced since the start of the project in early 2009. For example:

- Dense SNP chips as used in pig and sheep case studies were not available for use, although they were anticipated by the project proposal;
- Pig and sheep reference sequences were not published;
- The 800k dense SNP chip used in this project for cattle was not formally planned, and there are now chips of this size available from both Illumina and Affymetrix;
- Sequence capture and exon capture kits were not available for routine use;
- After its introduction, the cost of sequence capture halved during the course of the project.

It is to be expected that developments will continue at a rapid pace:

- The density of SNP chips will increase making the identification of homozygous segments cleaner
- The information on sequence, and hence SNP order, will remove some of the existing pitfalls in identifying homozygous segments
- The process of sequencing, whether for a segment of the genome or the whole genome will become more affordable, making the step from homozygous segment to causal mutation more routine
- The annotation of sequences will become more comprehensive and bioinformatic interpretation of sequence will become more reliable, in turn making the identification of putative causal mutations more reliable.

The techniques for assembling **haplotype libraries** from dense SNP chip data on a breed will improve leading to the development of libraries for each mainstream breed, albeit at differing rates depending on the breed. In turn, this will allow more informed assessments of the statistical significance of observing homozygous segments.

Annex 1:

Glossary

Allele. Alternative form of a gene; a single allele for each gene is inherited from each parent (e.g. at a gene for eye colour the allele might result in blue or brown eyes).

Biobank. A collection of biological material and the associated data and information for a population stored in an organised system.

Chromosome. A chromosome is a discrete block of DNA and is one of the basic structures of the genome. All nuclear DNA is organised into chromosomes with the number varying between animal species. Genes on a chromosome are linked and tend to be inherited together.

Controls. Here controls is used a general term for those animals that are not cases i.e. are assumed not to exhibit the disease.

Dense SNP chips. A major breakthrough in the use of DNA technology was the creation of dense SNP chips. Genotypes for SNPs are easily and cheaply obtained. Moreover many thousands (or millions) of SNPs can be assayed simultaneously on a small pre-prepared plate called a chip. This makes it feasible to obtain thousands (or millions) of genotypes for a single individual at an affordable price. Also see 'SNP'.

Dominant. When a gene displays dominance the same phenotype is seen whether the dominant allele is homozygous or heterozygous. Also see 'Recessive'.

Epidemiology. The study of the cause, effects, incidence and distribution of disease in populations.

Exon. Not all of the genome codes for proteins, since it also contains information on how and when to build proteins. Those sections of the genome that do code for proteins are called exons.

Gene. The basic unit of heredity. Each gene has two or more forms (alleles) which can be the same or different. The gene includes those segments of DNA that precede and follow the sequence that codes for a protein (exons), as well as intervening sequences (introns).

Genome. All the genetic material in the chromosomes of a particular species or organism.

Genotype. The genetic material inherited from parents (not all of it is necessarily expressed in the individual).

Haplotype. A haplotype is a combination of alleles over (closely) linked genes or markers carried on a single chromosome. Haplotypes therefore tend to be inherited as a unit, but change over generations by recombination. Also see 'Recombination'.

Haplotype Library. A haplotype library for a population has the objective of listing all the haplotypes that may be found in a population for every region of the genome.

Annex 1

Homozygote. A homozygote is an individual carrying two similar alleles of a gene, e.g. QQ or qq. Homozygous is the adjective describing a homozygote for a gene. Homozygosity is a general term for the state. Also see 'Heterozygote'.

Linkage. Linkage is the phenomenon by which genes that are close together on a chromosome and which have been inherited together from one parent of an individual tend to be passed on together to an individual's offspring. The closer the genes are on a chromosome the stronger this phenomenon. When the loci are on different chromosomes this tendency is completely absent.

Linkage disequilibrium. Linkage disequilibrium is the non-random association of alleles in haplotypes. Over time recombination events between genes will remove this association. The further away the genes are from each other on the chromosome, the more quickly recombination events will remove the association. Also see 'Recombination'.

Locus. A locus is a position in the genome i.e. a position, such as the specific location of a gene, on a chromosome. The plural is loci.

Nucleotide. DNA in the nucleus has 4 different forms, called bases, denoted A, C, G and T. The order of these bases in groups of 3 is the genetic code mapping sequence to the different types of amino acid which build proteins.

Quantitative Trait Loci (QTL). Loci that affect quantitative traits (in which the phenotypes show continuous (numerical) expression, i.e. expressed in a continuously variable manner as opposed to "present/absent"). Often synonymous with genetic markers, but QTL may also be genes of known effect.

Recessive. A recessive allele is an allele that only has an effect on the phenotype when it is homozygous. Therefore if allele q is recessive, qq yields a different phenotype from Qq and QQ, which have the same phenotype. Q is said to be the dominant allele. It is an example of non-additive gene action. Also see 'Dominant'.

Recombination. The breaking and rejoining of DNA strands to form new molecules of DNA encoding a novel set of genetic information.

Reference Sequence. Once one animal of a species has been sequenced, it is less technically challenging to sequence others of the species, since the initial sequence is used as a reference for interpreting the technical data obtained from other animals. Therefore the first sequence is called the reference sequence.

SNP. SNP is a ubiquitous acronym for a Single Nucleotide Polymorphism caused by a mutation at a single nucleotide (in contrast to a deletion or other mutational event). Hence the position in the genome is called a SNP locus. Typically a SNP locus has two alleles but it can have 4, with each of the 4 bases being found at that position. A typical livestock genome may contain upwards of 40 million SNPs. Also see 'Dense SNP chips'.

Standard Error. The standard error (s.e.) is a measure of confidence in the estimate of a parameter. As a rule of thumb the true value of the parameter has a probability of ~65% of lying within 1 s.e. of the estimate and ~95% within 2 s.e.s.

Annex 2:

Case Studies

The methods described in the RIDGENE Manual have been tested in a number of real world populations. The species and type of disease investigated together with the current status of the work (at time of going to print) is currently as follows:

Pigs:

- Neonatal leg weakness – a marker has been developed to allow informed breeding decisions to be made to reduce the disease incidence

Sheep:

- Dwarfism - a candidate mutation has been identified and is being tested
- Wobbler - initial work has suggested that the inheritance pattern is more complex than that expected for a recessive inherited disease - therefore further cases are being examined.
- Hydrocephaly - available samples provided insufficient DNA quantity for the technique to be applied, therefore further samples are required.

Cattle:

- Knuckled hocks - Initial investigations have suggested that the inheritance pattern could be typical of that expected for a recessive inherited disease. Genomic investigations are still in the early stages and no conclusions have yet been drawn.
- Protruding tongue - Initial investigations have suggested that the inheritance pattern could be typical of that expected for a recessive inherited disease. Genomic investigations are still in the early stages and no conclusions have yet been drawn.
- Cardiomyopathy - initial work has suggested that the inheritance pattern is more complex than that expected for a recessive inherited disease - therefore further cases are being examined.

Knowledge Transfer Network

Biosciences

Biosciences KTN is Sponsored by the Technology Strategy Board, the Scottish Government, European Regional Development Funds, Defra, BBSRC and NERC.

Technology Strategy Board
Driving Innovation

The Scottish Government

defra
Department for Environment Food and Rural Affairs

BBSRC
bioscience for the future

NATURAL ENVIRONMENT RESEARCH COUNCIL

EUROPE & SCOTLAND
European Regional Development Fund
Investing in your Future

Contact us

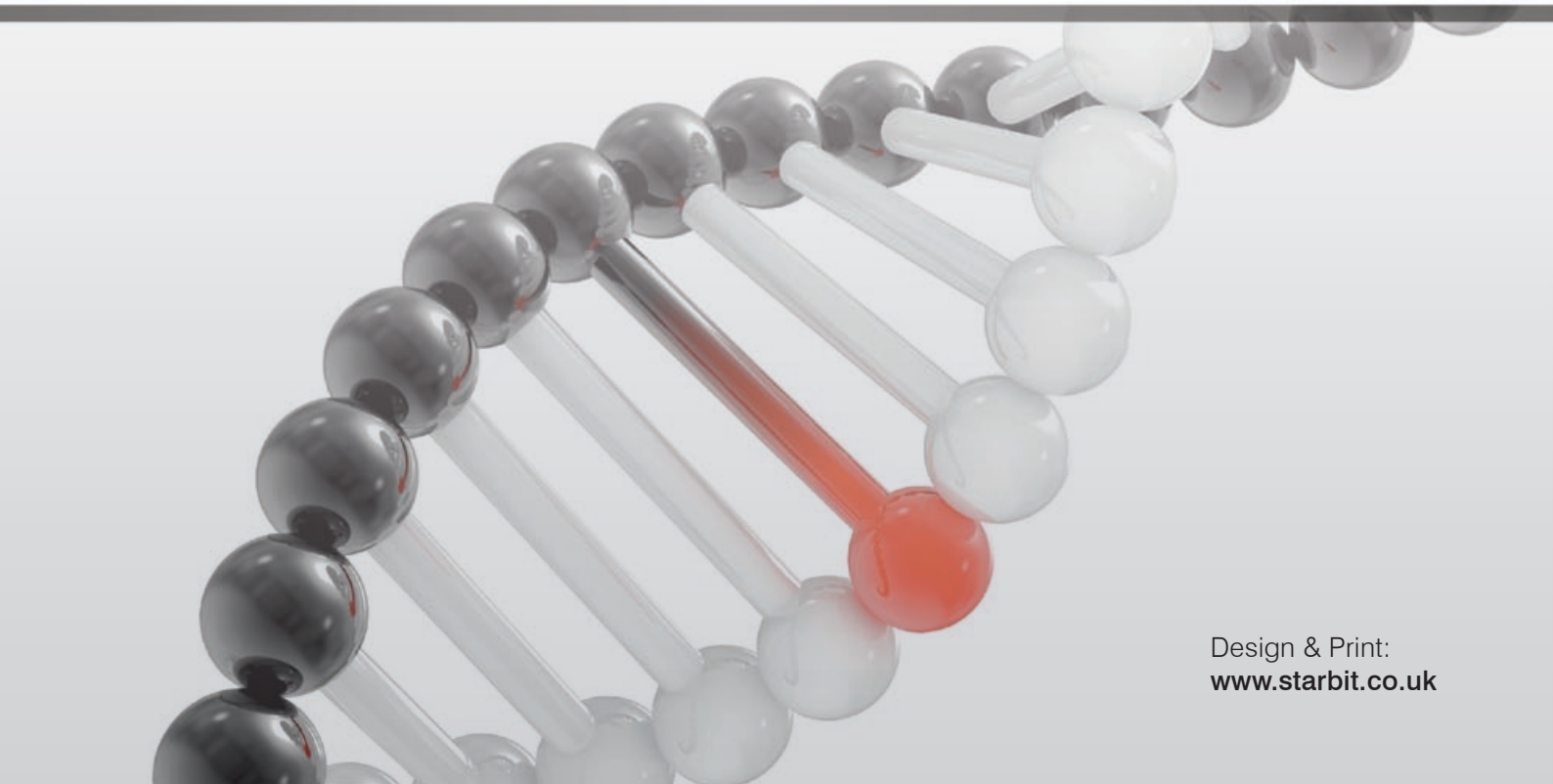
Biosciences KTN,
The Roslin Institute,
Easter Bush, Midlothian.
EH25 9RG.

Tel: +44 (0)131 651 7334
Fax: +44 (0)131 651 7335
Email: info@biosciencektn.com

Get _connected

www.innovateuk.org/biosciencesktn

Register now to stay up-to-date with all the latest developments



Design & Print:
www.starbit.co.uk