COMPLEXITIES AND STRATEGIES TO IDENTIFY THE CAUSATIVE MUTATION RESPONSIBLE FOR SINGLE LOCUS INHERITED DISEASES IN LIVESTOCK.

J. A. L. Cavanagh, I. Tammen and H. W. Raadsma.

Reprogen, Faculty of Veterinary Science, The University of Sydney, Camden, NSW 2570

SUMMARY

With the advent of high density SNP genotyping chips for livestock species, it is possible to quickly identify genomic regions of interest for well characterized inherited diseases given appropriate sample material. However, the search for causative mutations may be severely hampered if there is no functional candidate gene in the region, if the candidate genes are large, or if the disease causing mutation is not located in coding regions. Further complications are likely if a reference genome sequence is not available or incomplete. A quick scan of the coding region of a small number of candidate genes using traditional Sanger sequencing techniques can yield rapid results in amino acid changing mutations – either on genomic DNA or cDNA. If no coding mutations are found, a *de novo* search for disease causing mutations in genomic regions of interest is required. The use of next generation sequencing technologies in combination with targeted sequence capture is a recent development which so far eluded systematic searches studying the genetic architecture for simple monogenic and possibly complex traits.

INTRODUCTION

There is a worldwide interest in the identification of genes and causative mutations in those genes that cause single locus inherited diseases in livestock and companion animals. Whole genome sequencing has progressed rapidly for a number of livestock species and companion animals. From the whole genome sequencing effort, high density SNP genotyping arrays (chips) have been produced with unparalleled capacity to screen entire genomes with relative ease.

Inherited diseases are often disseminated widely with the use of popular sires. The aim is to characterize the mutation causing the defect and develop a direct DNA-test available for use by livestock industries to reduce the frequency of the disease allele and therefore the incidence of affected animals in the population. Direct and immediate benefits of a DNA-test are the reduction and minimization of losses caused by the defect and improvement of animal welfare. Indirect benefits pertain to Australia's reputation as a producer of high quality livestock with minimal incidence of major disease and genetic defects. In addition, the study of genetic defects may have medical significance as a large animal model for a homologous human disease.

There is an expectation that these genes and corresponding causative mutations would be relatively straightforward to map given the simple mode of inheritance, clear penetrance and a well defined phenotype. This is not always the case, although over the last few years, we have successfully identified causative mutations for Dexter cattle chondrodysplasia, Neuronal Ceroid Lipofucinosis (NCL) in Devon cattle and Merino sheep, and zinc deficiency in Angus cattle. We are currently implementing new strategies as discussed in this paper to search for causative mutations for NCL in South Hampshire sheep, Fawn Calf Syndrome (FCS) in Angus cattle and several other emerging defects in livestock. The challenge for all such diseases has not been the mapping or genome localization but rather the fine search for causative mutations underlying the phenotype. The advent of new large scale sequencing and screening technologies is likely to revolutionize the rate at which causative mutations are identified despite some pitfalls.

STRATEGY AND POTENTIAL PITFALLS

In the past, large scale microsatellite screening projects were employed to identify an area of the genome in linkage with the specific disease or phenotype of interest if no obvious candidate gene was identified. This was a time consuming and costly approach but relatively powerful given the specific location of a causative underlying genetic variant and a clear phenotype. Furthermore, in the case of recessive disorders, regions of homozygosity were expected within LD around the causative gene, thus allowing large scale comparison of affected, obligate carrier, and control animals. However, fine mapping to a single locus region remained problematic and in the absence of a suitable positional candidate gene, a challenge to map down to a single mutation. If functional candidate genes have been identified, homozygosity mapping can proceed in the targeted regions containing such candidate genes (Tammen *et al.* 1999). In one such case, the identification of the gene causing chondrodyplasia in Dexter cattle (Cavanagh *et al.* 2007), 11 functional candidate genes were selected and microsatellite markers targeted to these regions. Fortunately in this case, by way of comparative analysis across species, one of these candidate genes was later shown to be the target gene responsible, but in many cases the area of linkage identified is large (e.g. >3MB), spanning many genes and the causative mutation is not found. With the advent of SNP chips, identification of a region of interest has been greatly simplified. By sending away a sufficient number of informative DNA samples to a service laboratory (e.g Illumina bovine 54,000 SNP chip), a target region can be readily identified using a homozygosity mapping approach. A region of homozygosity amongst affected animals is sought whilst carrier and control animals are heterozygous (or not homozygous for the allele linked to the disease phenotype). The region of homozygosity can potentially be reduced by increasing the number of samples analysed and therefore mapping historical recombination events. The SNP chip strategy was employed in several studies (Charlier *et al.* 2008) as well as for mapping the homozygous region of FCS in Angus cattle to a region of 3.5Mb, which was further reduced with additional targeted SNP typing (unpublished data).

On identifying a genome region of interest a well annotated genome sequence either in the species of interest or via comparative approaches allows for identification of any functional candidate genes. Those diseases which are characterized biologically as phenotypes, enable similar diseases to be identified in human and mice, often with known causative genes. Additionally, the use of mice knockout models for each of the potential target genes, may allow for phenotype comparisons relevant to each gene. However, there are still many anonymous genes with unknown function, and it may not be possible to identify a likely candidate gene. One avenue to narrow down the list of relevant genes underlying a target region is through transcriptome analysis (gene expression studies) and looking for expression differences amongst affected and control animals. The usual constraints of timing of sampling and tissue specific expression prevent this approach from being definitive.

Traditional Sanger sequencing methods of positional candidate genes, can be used in amplified PCR products covering the coding region of a gene(s) of interest and sequenced for putative causative mutations. Again, this is time consuming and costly. This method was successful for several projects in our laboratory including Dexter cattle chondrodysplasia in which 2 causative mutations were identified in a single gene (Cavanagh *et al. 2007*), and single coding mutations in NCL in Devon cattle (Houweling *et al.* 2006), NCL in Merino sheep (Tammen *et al.* 2006), and Zinc deficiency in cattle (Tammen *et al.* 2002). All of these mutations were found in exons (coding and non-coding). With strong evidence for a positional candidate and high likelihood of an exonic mutation, this is still a preferred method of choice to map a mutation relatively quickly. However in the absence of a strong candidate gene, and a large region of homozygosity harbouring multiple genes, sequencing of all genes in a 3MB region has proven prohibitive until recently. With the advent of next generation sequencing technologies such as Roche 454, Illumina Solexa,

and AB SOLiD, a large volume (up to multiple Gigabases/sample) of sequencing data is obtainable. Each of these technologies have different strengths and weaknesses as discussed by Harismendy *et al.* (2009). There are several strategies such as whole genome re-sequencing, end sequencing of mate pair libraries, and sequence capture using arrays are of particular interest. To specifically target a region, a targeted DNA capture and sequence service, such as NimbleGen (http://www.nimblegen.com/products/seqcap/index.html), Febix (http://www.febit.com/go/en/services/hybselect/), Agilent (http://www.opengenomics.com/SureSelect) and LC Sciences (http://www.lcsciences.com/products/genomics/targeted_sequencing/targeted_sequencing.html) is required. In general the procedure involves fragmenting the DNA into small pieces and hybridizing it to a custom array containing probes to match the sequence of interest. The DNA of interest is then eluted and directly sequenced using a next generation sequencing platform. The major limitation is where no reference sequence is available to design capture probes. In this case a preliminary BAC sequencing step is required covering the target to generate denovo reference sequence.

DE NOVO MUTATION SCREENING/FINDING CAUSATIVE MUTATIONS

By implementing targeted sequence capture and next generation sequencing technology, the DNA sequence of control versus affected animals can be compared for differences. Not only does this reduce the emphasis on identification and characterization of positional candidate genes before hand, it also provides non-coding sequence information within the region of interest. In the case of FCS in Angus cattle and NCL in South Hampshire sheep, coding regions of each respective candidate gene were screened for mutations with a negative result. Now, with the use of targeted sequence capture and next generation sequencing, large amounts of DNA can be sequenced which will lead inevitably to the identification of many possible mutations, which in turn may lead to false positive direct DNA-tests. By carefully selecting animals to sequence, such as groups of full/half sibs (control/affected pairs), chance differences can be reduced. However, there will still be many mutations in perfect linkage disequilibrium with the disease and it may not yield direct information on the likely causative mutation. Furthermore, many elements in the genome may act as regulatory factors, copy number variants, repeat elements and changes in transcription factor binding sites or micro-RNAs. Such mutations are subtle and may not lead to an obvious causative role.

TOWARDS DEFINITIVE PROOF ON CAUSATIVE MUTATION

Simple single nucleotide polymorphisms (SNPs) can be causative mutations but difficult to substantiate if they are not directly related to predicted changes in the amino acid sequence. Mutations affecting coding sequence often mean that the RNA is subject to non-sense mediated RNA decay (Frischmeyer and Dietz 1999), resulting in little or nill of the mutant RNA being translated into protein. This is not the case for regulatory mutations that can affect the expression levels of the RNA and subsequent translated protein. Once a putative causative mutation has been identified, it is extremely difficult to prove without using allele substitution which may be done in appropriate cell lines, but is not generally done *in vivo*. There are many mouse knockout models which support a specific gene function, but these are generally not sensitive to define all possible mutations which could lead to the same phenotype and are therefore a blunt instrument. Targeted induction of the mutation in normal animals remains difficult but provides a powerful level of proof to a gold standard. Furthermore long range trans-acting regulatory elements, although rare, may also affect a phenotype and it will be difficult to identify these if not captured within the target sequencing strategy.

IMPLICATIONS FOR COMPLEX POLYGENIC TRAITS

Although the search for underlying causative mutations or QTN in polygenic traits is of interest, such searches will be an order of magnitude more difficult. Firstly with any population the strong LD which exists between markers surrounding the QTN will be extensive with many ambiguous SNP as possible QTN. Furthermore in the absence of definite proof of a gene let alone a specific mutation in a complex pathway, the power to attribute phenotypic changes to genes of relatively small effect is problematic. This is further hampered where compensatory or alternative pathways may act on the same phenotype expression such as in disease /immune response. As seen in studies with human height, a quantitative polygenic trait with a heritability of >80%, simply increasing the density of SNP to ultra high densities is not sufficient to account for all the additive genetic variance. Large scale sequencing of individual genomes will become a reality within the near future, but will only add to the complexities of dealing with far more explanatory variables than observations and most studies will fail to definitively map all causative mutations.

CONCLUSIONS

With the advent of new technologies, high density SNP screening, whole genome re-sequencing and targeted sequence capture, the characterisation of inherited diseases in livestock will become easier. Studies can proceed directly from whole genome SNP genotyping to targeted SNP genotyping panels if required, then to targeted sequence capture and sequence thus avoiding the steps of PCR amplification and Sanger sequencing. Diseases that have not yet had a causative mutation identified can be processed through the targeted sequence capture process. This method is also extremely efficient for characterising new or emerging inherited diseases in livestock. However, proof to a gold standard that the predicted causative mutation identified remains extremely difficult. Once predicted causative mutations are validated they can be used as a direct DNA-test to identify carrier animals and eliminate the deleterious defect in livestock populations and may be assembled as a single test for routine applications.

REFERENCES

Cavanagh, J.A., Tammen, I., Windsor, P.A., Bateman, J.F., Savarirayan, R., Nicholas, F.W. and Raadsma, H.W. (2007) *Mamm Genome*. **18**:808

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. and Georges, M. (2008) *Nat Genet*. **40**:449

Frischmeyer, P.A., and Dietz, H.C. (1999) *Hum Mol Genet* **8**:1893

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009) *Genome Biol*. **10**:R32. [Epub ahead of print]

Houweling, P.J., Cavanagh, J.A., Palmer, D.N., Frugier, T., Mitchell, N.L., Windsor, P.A., Raadsma, H.W., and Tammen, I. (2006) *Biochim Biophys Acta*. **1762**:890

Tammen, I., Cavanagh, J.A.L., Harper, P.A.W., Cook, R.W., Raadsma, H.W., and Nicholas, F.W. (1999). *Archives of Animal Breeding.* **42**:163

Tammen, I., Cook, R.W., Gitschier, J., Nicholas, F.W., and Raadsma, H.W. (2002) *Proceedings of the 28th International Conference on Animal Genetics*. C059 p70

Tammen, I., Houweling, P.J., Frugier, T., Mitchell, N.L., Kay, G.W., Cavanagh, J.A., Cook, R.W., Raadsma, H.W., and Palmer, D.N. (2006) *Biochim Biophys Acta*. **1762**:898