Proc. Assoc. Advmt. Anim. Breed. Genet. 22:1-8

FUNCTIONAL ANNOTATION OF ANIMAL GENOMES

E.L. Clark, S. Bush, R. Young, J.K. Baillie, L. Lefevre, P. Dutta, C. Muriuki, M. McCulloch, T.C. Freeman, D.W. Burt, L. Freem, C.B.A.Whitelaw, K.M. Summers, A.L. Archibald & D.A. Hume

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

SUMMARY

With the advent of long-read sequencing technologies, and the rapid drop in the cost of short-read sequencing, livestock geneticists have access to almost completely contiguous reference genome sequences of similar quality to human and model organisms, and massive sequence level data on variation amongst breeds and adapted populations. In livestock genomes, many protein-coding genes are marked with placeholder names, their functional orthology to human or mouse genes is ambiguous and the annotation of transcript diversity is sparse. Non-coding regulatory elements (promoters, enhancers etc) and non-coding RNAs are even less well characterised, yet available evidence from human genetics indicates that variants in these elements are enriched for trait associations. The international FAANG (Functional Annotation of Animal Genomes, www.faang.org) consortium aims to coordinate efforts to address the information gap (L. Andersson *et al.* 2015). Gene-editing technologies, combined with sequence information, offers the promise of accelerated genetic gain (Hickey *et al.* 2016). In this review, we consider some of our approaches to livestock genome annotation.

INTRODUCTION

At the previous meeting of AAABG, Perez-Enciso et al. (2015) (Perez-Encisco *et al.* 2015) reviewed the potential applications of sequence data to animal breeding; and talked of "biology-informed sequence exploitation". Since 2015, the cost of generating whole genome shotgun sequence data has continued to fall. Thus, with the most recent genotyping platforms, the $1000 genome at 30X genome coverage is not far from reality, and we and others are sequencing hundreds, and even thousands, of animals from different breeds and different adapted populations in every livestock species. The increased sequence depth increases the reliability of variant calling, including variants that impact on the function of protein-coding genes such as indels, stop gains and severely disruptive mutations (Boschiero *et al.* 2015, Telenti *et al.* 2016). These mutations are more prevalent in populations than might be expected. In a remarkable study of human populations with high levels of consanguinity, Saleheen et al. (Saleheen *et al.* 2017) reported exome sequencing of >10,000 individuals, and identified 49,000 rare predicted loss-of-function mutations of which 1317 were homozygous in at least one individual. A subset was confirmed to cause functional changes in the encoded protein, albeit clearly not lethal. An exome sequencing platform has been developed for pigs, and its application similarly predicts significant prevalence of loss-of-function alleles (Robert *et al.* 2014). This is a potential resource for functional genomics, as well as animal breeding, since the impact of such alleles can be confirmed by brother-sister mating or from prohibited homozygosity in populations (if the impact is severe). We have initiated such as backcross project in chickens, where we identified candidate loss-of-function alleles in a set of 10 founder pairs, and then mated their F1 offspring to expose homozygotes. However, even high impact functional variants are not necessarily coding. Hoff et al. (Hoff *et al.* 2017) identified seven haplotypes that were relatively prevalent in registered US Angus cattle, but were not observed as homozygotes, and used deep sequencing of >100 individuals to identify common variants within these haplotypes. None of the candidate causal variants identified was present within exons.

Another of the major impacts of deep sequencing is the improved detection of copy-number variants and sequences that are not present in the reference genome. This is somewhat constrained by the quality of the genome assembly (Couldrey *et al.* 2017) but the rapid improvement of livestock genomes, driven by the FAANG consortium, will address this issue. Indeed, the contiguity of the new goat genome, released earlier this year (Bickhart *et al.* 2017, Worley 2017), is approaching that of the completed human and mouse genomes. The recent sequencing of 10,000 human genomes at 30-40x coverage identified on average 0.7 Mb of sequence that was not present in the human reference genome (Telenti et al. 2016). Copy number and structural (e.g. inversions/translocations) variants are commonly associated with trait variation in all species. A recent study, which also reviewed some of the earlier literature, identified multiple copy number variants associated with domestication and high altitude adaptation in the Chinese Yak (Zhang *et al.* 2016). Long read sequencing provides an additional potential step-change in detection of structural variants, with an incomplete overlap between the outcomes from short-read technologies (Couldrey et al. 2017). With all of this sequence/genomic information, we have the potential to reverse the traditional information flow, and link sequence to consequence. However, there are several major challenges to overcome.

Firstly, we need much more information about the function of individual genes and regulatory sequences in a wider range of species. It is certainly the case that some functions are conserved across species. The phenotypes associated with knockouts of protein-coding genes in mice can give insights into likely functions and phenotypic consequences of loss-of-function in other species. Similarly, detailed analysis of promoter and enhancer landscapes in the liver across 20 mammalian species revealed substantial conservation of both regulatory elements and transcriptional outputs (Villar *et al.* 2015). Arguably, the liver has a rather generic "housekeeping" function in mammals that is not subject to rigorous selection. By contrast, there are radical differences between mice, pigs and humans in the response of innate immune cells to bacterial lipopolysaccharide (LPS) (Kapetanovic *et al.* 2012, Schroder *et al.* 2012) or to glucocorticoids (Jubb *et al.* 2015), associated with gain and loss of promoter and enhancer elements. It is these differences between species, and between individuals, that are of particular interest to geneticists and developmental biologists.

Secondly, we need to find a way to take account of epistasis, which manifests as variable penetrance. There are few knockout mutations in mice, or human genetic diseases, that do not exhibit some measure of phenotypic variation that is apparently a consequence of gene-gene interactions, or genetic background (Phillips 2008, Mackay 2014). Sometimes the mechanism can be disentangled based upon biological knowledge. For example, the knockout of the macrophage-specific transcription factor, PU.1, is mid-gestation lethal in homozygous PU.1 knockout inbred C57Bl/6 mice, but when the knockout allele is present in the homozygous state on a different genetic background, produces viable offspring with a neutrophil deficiency. The PU.1 protein interacts with another transcription factor, MITF, and a compound heterozygote (PU.1 +/-, MITF mi/+) phenocopies the PU.1 knockout (Luchin *et al.* 2001). Efforts to model the impact of epistasis in GWAS analysis and genomic selection have had limited success, in part due to the computational challenges (Stanislas *et al.* 2017). A subset of variable penetrance results from genomic imprinting in mammals, where the apparent heritability of a trait depends upon the parent of origin and reciprocal crosses do not produce the same outcome. The analysis of the contribution of imprinting to estimated breeding values is also computationally challenging (Nishio and Satoh 2015), but would be significantly less so if the set of imprinted loci and their functions was known in each species.

Identification of causal variants has been described as the "holy grail" for quantitative genetics (Perez-Encisco et al. 2015). Increased density of markers derived from sequence information, without functional annotation, simply approaches the tyranny of statistics. The challenge is to develop strong biological "priors" to prioritise variants that are more likely to be functionally associated with a trait. Inclusion of such biological priors clearly has the potential to enhance the power of genomic prediction

in complex traits (MacLeod *et al.* 2016). So, how far have we come since 2015 in generating useful prior knowledge?

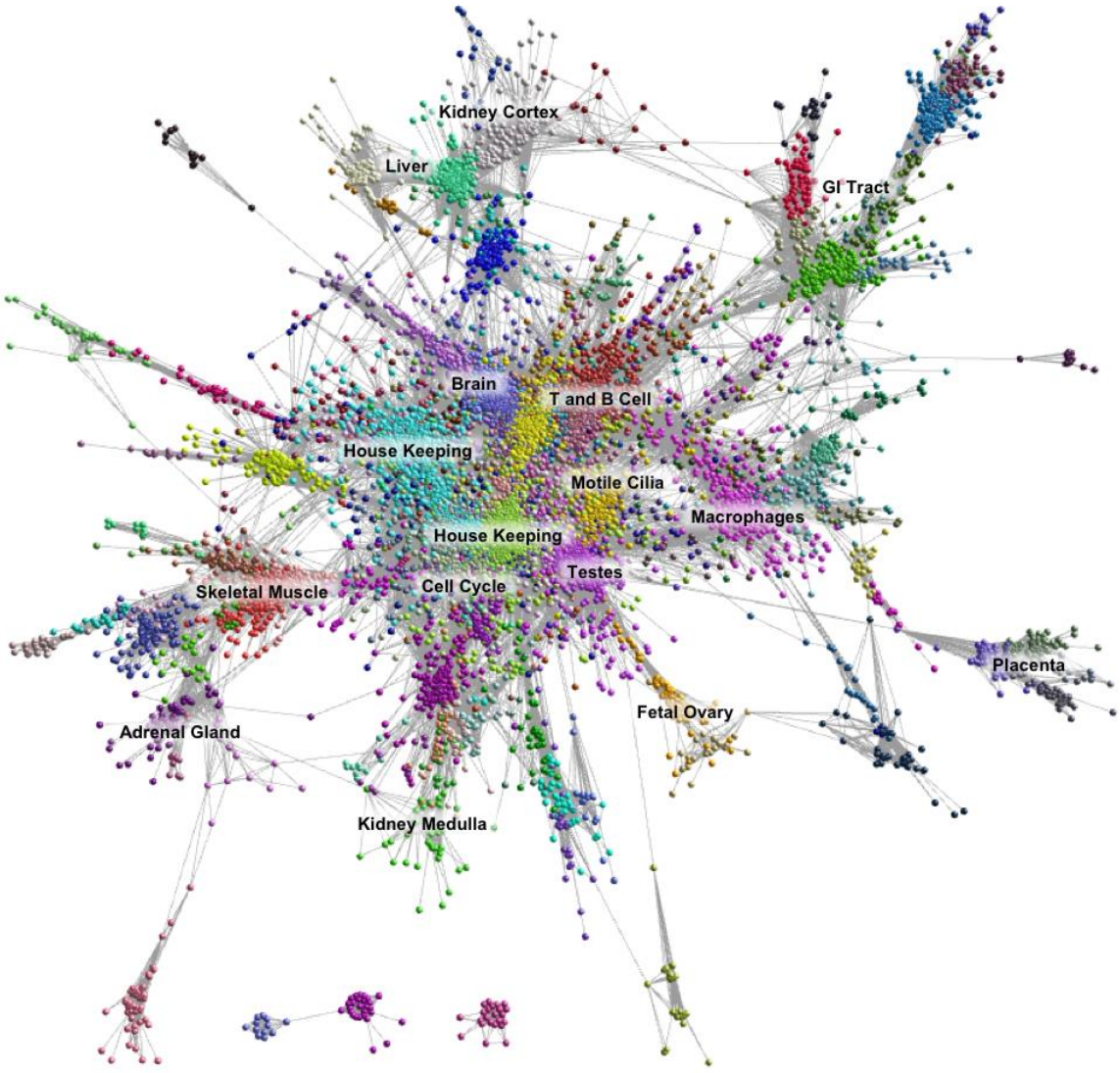

Figure 1. The transcriptional network of the sheep gene expression atlas dataset. Each node represents a single transcript, the lines between them represent correlations (edges) and the colours are shared by nodes that have correlated expression across the network (The graph is comprised of 15,192 nodes (genes) and 811,213 edges, r = 0.75, MCLi = 2.2).

TRANSCRIPTIONAL ATLAS PROJECTS

All of the processes that underpin development, growth, physiology and productivity depend upon the functions of numerous gene products that act together to generate pathways, macromolecular complexes, organelles, cells, organs and systems The set of genes required to deliver a cell-type, an organelle or a functional complex must share transcriptional regulation, so that their products are available in the correct place at the right time. If one samples the transcriptome of many different organ and cellular systems that differ from each other, the levels of transcripts encoding products that function together must be correlated with each other. The more physiological states that one samples, the more stringently one can determine that a pair of genes shares strict coexpression. Since the pioneering efforts that produced the Symatlas (now BioGPS, http://biogps.org) from sets of microarray data from mouse and human cells and tissues, there has been an explosion of gene expression "atlases" across multiple tissues in a number of species and within tissues across cell types and developmental time in humans and mice. The principal of guilt by association, namely that one can infer a great deal about the likely function of a gene product from its transcriptional neighbours, was clearly fulfilled in analysis of the mouse BioGPS dataset (Hume *et al.* 2010). For example, the entire set of genes encoding the lysosome was co-expressed, and specifically elevated in phagocytes. Similarly, genes involved in the cell cycle, in protein synthesis, or in extracellular matrix, clearly formed co-expression clusters because they are regulated activities and different cells and tissues engage these pathways to different extents. The exception is the set of genes that is relatively ubiquitously-expressed: the house-keeping genes. The housekeeping gene set also contains the highest proportion of genes that lack informative annotation, a reflection of the focus of biologists on differential expression. To identify and visualise transcriptional clusters in very large datasets, we utilized the network-clustering tool Biolayout *Express*^{3D}, now developed as Miru (http://www.kajeka.com). One advantage of the consistency of commercial microarray platforms was that it was possible to consolidate and integrate data from multiple laboratories, for example to generate an atlas of gene expression in human cells (Mabbott *et al.* 2013), also available as a default set on BioGPS.

The generation of transcriptional atlases for livestock species is more recent. We utilized extensive EST data to design a comprehensive microarray for the pig, and created a transcriptional atlas (Freeman *et al.* 2012). One example of the principal of guilt-by-association was the identification of a comprehensive set of transcripts associated with mitochondrial oxidative phosphorylation, and separation of the nuclear and mitochondrial-encoded transcripts (indicating that their transcription is not perfectly correlated). A bovine expression atlas was generated based upon tag sequencing of tissue from adult, juvenile and fetal tissues (Harhay *et al.* 2010) and subsequently extended in a set of 18 tissues from a single animal by RNAseq (Chamberlain *et al.* 2015). More recently, we have produced an extensive transcriptional atlas based upon direct sequencing of mRNA from six adult sheep as well as embryos and juveniles at various developmental ages (bioRxiv132696). The animals were deliberately chosen as cross breeds between the reference Texel (Jiang *et al.* 2014) and the Scottish Blackface. Figure 1 shows the overview of the transcriptional network, which clearly segregates the transcripts into tissue, cell-type and process-specific clusters. The latter clusters include a comprehensive set of genes involved in the cell cycle, protein synthesis, oxidative phosphorylation and motile cilia. Note also the close proximity of liver and kidney cortex in the network, indicating their similar expression profiles. We identified many transcripts encoding enzymes associated with gluconeogenesis and amino acid metabolism that are shared between the two organs. These data have also been made available on Biogps (biogps.org/sheepatlas). We are also currently analyzing similar projects, albeit on a smaller scale (guided by transcript diversity observed in the sheep) in commercial cross-bred goats, Indian and Mediterranean (the reference breed for the current assembly of a water buffalo genome) water buffalo and broiler and layer chickens.

These data together will produce a quantum leap in the analysis of transcript variants in each of the species and have contributed to the various genome projects to support improved annotation.

The next phase of genome/transcriptome annotation is the identification of regulatory elements. Several of the authors have had a long-term association with the FANTOM Consortium. The consortium utilized Cap Analysis of Gene Expression (CAGE) to generate a promoter-based atlas of gene expression in humans and mice (Consortium *et al.* 2014). CAGE, which is essentially genome-scale 5'RACE, also detects the short transcripts that are produced by active enhancers (R. Andersson *et al.* 2014) and the integration of information derived from detected promoter and enhancer activity can be used to infer the relationship between the two. Enhancers and promoters generated by CAGE sequencing were strongly correlated with similar elements detected by ChIP-seq analysis of the location of acetylated and methylated histones including data from the ENCODE consortium. In the analysis of a diversity of time courses of cell activation or differentiation, the transcriptional activity of enhancers in the vicinity of inducible genes was increased transiently in advance of detectable promoter activation (Arner *et al.* 2015, Baillie *et al.* 2017). The most recent FANTOM publication integrated CAGE and RNAseq data to identify 27,000 long non-coding RNAs encoded by the human genome, and to demonstrate that these transcripts derive primarily from enhancers. They further demonstrated that the lncRNAs that overlap trait-associated SNPs are expressed in cell types that are relevant to the trait in humans. The RNAseq data we have obtained from livestock species also greatly expands the diversity of lncRNAs identified and by inference, will contribute to the location of likely trait-associated regulatory elements. The FANTOM5 data from humans and mice can be usefully mapped across to other large animals such as pigs to identify conserved promoters and enhancers (Robert *et al.* 2015), in the process supporting other evidence that the transcriptome of pigs is substantially more human-like than that of mice.

APPLICATIONS OF TRANSCRIPTOMIC DATA IN GENETICS

SNPs associated with enhancers and promoters detected by the FANTOM5 consortium were more likely even than exonic SNPs to be associated with human disease susceptibilities (R. Andersson et al. 2014), mirroring evidence based upon identification of open chromatin detected as DNase1 hypersensitive sites (Maurano *et al.* 2012). More recently, genome-wide analysis of long range interactions between distal enhancers and promoters in multiple human cell types provided further links between regulatory variants and disease susceptibility traits (Javierre *et al.* 2016). The principle can be extended further. Regulatory variation in sets of genes that each contribute independently to a common pathway are likely to each contribute to a trait that depends upon that pathway. Consistent with the proposal, it is possible to identify and quantify co-expression of RNAs from trait-associated regions (bioRxiv, 095349) and from that information, to draw inferences about the likely underlying biology and to identify additional candidate susceptibility loci. Based upon that principle, we formed the hypothesis that genes involved in susceptibility to inflammatory bowel disease (IBD) were co-expressed specifically in monocytes and regulated during their differentiation. We identified a set of promoters that fulfilled that criterion and which were strongly enriched for associations with IBD, including >100 novel loci (Baillie et al. 2017).

The link between SNPs in regulatory regions and complex traits, of course has an intermediate phenotype in the form of heritable variation in the level of the regulated transcript, so-called expression quantitative trait loci (eQTL). Variation within such loci may act in cis or trans to produce differences in transcript abundance. Most evidence of eQTL to date has relied on microarray profiling of the same tissue or cell type from large numbers of individuals and conventional GWAS, or in defined crosses, an approach that has been called "genetical genomics" (de Koning *et al.* 2007, Martinez-Montes *et al.* 2017). Studies of human leukocytes have revealed that the large majority of

transcripts detected on a microarray display detectable and heritable variation in expression (Fairfax *et al.* 2014, Westra *et al.* 2015).

Sequence-based analysis of the expression of each allele in individual animals has the potential to massively increase the power of detection of eQTL (Almlof *et al.* 2012), and this approach has become substantially more straightforward with the feasibility of obtaining high depth coverage of DNA and RNA sequences from the same animal(s). Chamberlain et al (Chamberlain et al. 2015) utilized RNAseq data to demonstrate the pervasive allele-specific expression of genes in 18 tissues of a single cow, including a surprising level of mono-allelic or parent of origin-specific expression and tissue-specificity. The sheep genome consortium also noted pervasive mono-allelic expression in transcriptome analysis of the pure-bred Texels (Jiang et al. 2014). In our own RNAseq data from multiple species, we have deliberately chosen to analyse cross-bred animals, and sequenced a wider diversity of tissues at greater depth than previous studies. One of the advantages of deep sequencing is that unprocessed nuclear RNAs, and lncRNA are covered at sufficient depth to detect variation in expression, and these non-coding regions have much higher density of SNVs (Barreiro *et al.* 2008). The MBASED algorithm (Mayba *et al.* 2014) can be used to integrate expression estimates from multiple SNV level RNAseq counts, to integrate allele specific expression (ASE) detection across a locus. With sufficient sequencing depth, the analysis can extend into neighbouring regulatory regions without the requirement for phasing information. We are currently utilizing this approach to identify ASE in sheep, water buffalo, pig and chicken RNAseq datasets.

One of the applications of particular interest is to begin to understand the benefits of cross-breeding or heterosis. The molecular basis for the benefits of cross-breeding is relatively poorly understood, and much of the analysis comes from plants, rather than animals (Chen 2013). In the sheep transcriptional atlas, we were able to integrate data from a smaller RNA-seq atlas derived from pure-bred Texels, produced in association with the release of the sheep genome (Jiang et al. 2014). A subset of transcripts was much more highly-expressed in the muscle and brain in the cross-bred animals than in the pure Texel animals. If most trait variation is associated with transcriptional regulation, heterosis presumably derives from some form of optimal contribution of the variant expression alleles of each parent within the cell and tissues that control the trait. Combining data from transcriptional networks and allele-specific transcription in cross-bred animals may eventually underpin the prediction of cross-bred animal performance.

GENOME EDITING

Alongside the revolution in genome sequencing, genome editing technologies provide a second revolution; the capacity to confirm predictive functions by altering the genome in model organism or in the species of interest. However, genome editing is more likely to be deployed in farmed animal species to modify or delete protein coding genes in order to generate animals with desirable genotypes that cannot readily be established by conventional selective breeding. A couple of recent examples of such desirable traits are resistance to Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) and germline ablated male pigs that can serve as vehicles to increase the delivery of gametes from elite males (Burkard *et al.* 2017, Park *et al.* 2017). The use of primordial germ cells has expedited the application of germ-editing in poultry (Taylor *et al.* 2017). Perhaps more challenging is the prospect of accelerating genetic gain in breeding programmes by multiplex editing of functional variants in a single generation (Hickey et al. 2016), or even the application of so-called "gene drives" (Gonen *et al.* 2017). That prospect is certainly on the horizon, but the consequences of editing enhancer elements in mice have not been entirely predictable. Most genomic loci contain numerous apparently conserved and functional enhancers, and many others that are gained and lost between species (Villar et al. 2015). There is still some way to go before we can predict consequence from sequence in regulatory elements.

CONCLUSIONS

The availability of high throughput sequencing and its decreasing cost combined with development of new methods for modifying animal genomes has opened a wide range of approaches that will enhance genome annotation in livestock animals and lead to greater understanding of important production traits and processes such as heterosis.

REFERENCES

Almlof, J. C., et al. (2012). PLoS One **7**(12): e52260.
Andersson, L., et al. (2015). Genome Biol **16**: 57.
Andersson, R., et al. (2014). Nature **507**(7493): 455-461.
Arner, E., et al. (2015). Science **347**(6225): 1010-1014.
Baillie, J. K., et al. (2017). PLoS Genet **13**(3): e1006641.
Barreiro, L. B., et al. (2008). Nat Genet **40**(3): 340-345.
Bickhart, D. M., et al. (2017). Nat Genet **49**(4): 643-650.
Boschiero, C., et al. (2015). BMC Genomics **16**: 562.
Burkard, C., et al. (2017). PLoS Pathog **13**(2): e1006206.
Chamberlain, A. J., et al. (2015). BMC Genomics **16**: 993.
Chen, Z. J. (2013). Nat Rev Genet **14**(7): 471-482.
Couldrey, C., et al. (2017). J Dairy Sci.
de Koning, D. J., et al. (2007). Poult Sci **86**(7): 1501-1509.
Fairfax, B. P., et al. (2014). Science **343**(6175): 1246949.
FANTOM Consortium (2014). Nature **507**(7493): 462-470.
Freeman, T. C., et al. (2012). BMC Biol **10**: 90.
Gonen, S., et al. (2017). Genet Sel Evol **49**(1): 3.
Harhay, G. P., et al. (2010). Genome Biol **11**(10): R102.
Hickey, J. M., et al. (2016). J Anim Breed Genet **133**(2): 83-84.
Hoff, J., et al. (2017). BioRxiv.
Hume, D. A., et al. (2010). Genomics **95**(6): 328-338.
Javierre, B. M., et al. (2016). Cell **167**(5): 1369-1384 e1319.
Jiang, Y., et al. (2014). Science **344**(6188): 1168-1173.
Jubb, A., et al. (2015). Lancet **385 Suppl 1**: S54.
Kapetanovic, R., et al. (2012). J Immunol **188**(7): 3382-3394.
Luchin, A., et al. (2001). J Biol Chem **276**(39): 36703-36710.
Mabbott, N. A., et al. (2013). BMC Genomics **14**: 632.
Mackay, T. F. (2014). Nat Rev Genet **15**(1): 22-33.
MacLeod, I. M., et al. (2016). BMC Genomics **17**: 144.
Martinez-Montes, A. M., et al. (2017). Mamm Genome **28**(3-4): 130-142.
Maurano, M. T., et al. (2012). Science **337**(6099): 1190-1195.
Mayba, O., et al. (2014). Genome Biol **15**(8): 405.
Nishio, M. and M. Satoh (2015). Genet Sel Evol **47**: 32.
Park, K. E., et al. (2017). Sci Rep **7**: 40176.
Perez-Encisco, M., et al. (2015). Association for the Advancement of Animal Breeding Proceedings 21: 21126.
Phillips, P. C. (2008). Nat Rev Genet **9**(11): 855-867.
Robert, C., et al. (2014). BMC Genomics **15**: 550.
Robert, C., et al. (2015). BMC Genomics **16**: 970.
Saleheen, D., et al. (2017). Nature **544**(7649): 235-239.
Schroder, K., et al. (2012). Proc Natl Acad Sci U S A **109**(16): E944-953.

Stanislas, V., et al. (2017). BMC Bioinformatics **18**(1): 54.
Taylor, L., et al. (2017). Development **144**(5): 928-934.
Telenti, A., et al. (2016). Proc Natl Acad Sci U S A **113**(42): 11901-11906.
Villar, D., et al. (2015). Cell **160**(3): 554-566.
Westra, H. J., et al. (2015). PLoS Genet **11**(5): e1005223.
Worley, K. C. (2017). Nat Genet **49**(4): 485-486.
Zhang, X., et al. (2016). BMC Genomics **17**: 379.