Looking back to move forward – a personal perspective on pig molecular genetics from RFLPs to nextgen sequencing

Christopher Moran

Faculty of Veterinary Science, University of Sydney, NSW, 2006.

Introduction

My first foray into porcine molecular genetics commenced about 27 years ago with a small Australian Research Council grant to develop and exploit DNA markers in the pig for genetic mapping. At that stage, the standard procedure for detection of DNA polymorphism involved digestion of genomic DNA with one or more restriction enzymes, electrophoresis, transfer to a membrane, hybridisation with a radioactively labelled probe (using dangerous radioactive phosphorous) and autoradiographic detection on an X-ray film. The detection of these restriction fragment length polymorphisms (RFLPs) was slow, insensitive, expensive and hazardous, but was an important starting point and enabled initial molecular studies of porcine stress syndrome and early contributions to the international PiGMaP Consortium established in 1991 to develop a comprehensive linkage map for the pig. The Pig Research and Development Corporation, later to become Australian Pork Limited, soon began to fund research in my laboratory, continuing up until 2005, with continuous funding of a postdoctoral position from 1992, initially for Paul Le Tissier and subsequently for Yizhou Chen. This funding went on to enable a multi-institutional program of co-operative research involving Universities of Sydney, Melbourne and New England as well as the industry partner QAF (Bunge) Meat Industries in search of quantitative trait loci (QTL) for important traits involved in growth and productivity, meat quality and immune function.

Many important technical improvements to molecular biological techniques have occurred since then. These have included things as simple as DNA sample preparation, which originally involved cumbersome and expensive caesium chloride density gradient ultracentrifugation, now replaced by ion exchange columns or even simpler methods. The polymerase chain reaction (PCR) had a revolutionary impact enabling safer and much more effective detection of PCR-RFLPS and later the much more informative microsatellite genetic markers.

The PiGMaP Consortium formed the basis of an extremely useful international network that went on to develop plans for a genome sequence for the pig. Linkage and physical maps, particularly the very high resolution radiation hybrid maps, were important enabling resources for assembly of the porcine genome sequence, which generated using clone based sequencing and Sanger sequencing technology. An estimated $24.3 million from USA, Europe and China were spent on the development of this sequence with the first draft assembly released in 2009. The current Build10 of this genome sequence consists of over three billion nucleotides. Important by-products of the genome sequencing effort are the many hundreds of thousands of single nucleotide polymorphisms (SNPs), and chips for cheaply and effectively genotyping these polymorphisms en masse. These provide opportunities for very high resolution genetic mapping, including linkage disequilibrium mapping, for identifying genes (QTLs) involved in economically important traits, accurate retrospective reconstruction of relationship pedigrees and perhaps most significantly genomic selection.

It is instructive though that technical developments in sequencing, namely massively parallel sequencing, particularly the Illumina nextgen sequencing methodology, have so revolutionised genome sequencing that a complete mammalian genome sequence at high depth of coverage can be generated for about $10,000, with most of the difficulty and expense now involved in assembly and bioinformatic manipulation of the sequence. Near future applications of even more advanced sequencing methodologies will likely render redundant the genotyping of DNA polymorphisms even using highly cost effective SNP chips since the cost of generating complete but low coverage genome sequence will be trivially small and will potentially cover all variation, not just that located on existing chips.

From genetic map to genome sequence

From a virtually non-existent base in 1990, the genetic map of the pig has grown through numerous iterations of linkage (Archibald *et al*, 1995) and physical mapping to eventually deliver over 3 billion bases of ordered and annotated sequence in a complete genome assembly. Groenen, Schook and Archibald (2011) have documented the development of the genome sequence for the pig. While much still remains to be discovered and exploited from this sequence, and indeed all other genome sequences including the human, the generation and analysis of this sequence is an event of monumental significance. Table 1 summarises the content of this publicly available sequence which contains about 21000 protein coding genes and close to 3000 non RNA genes that don’t encode protein.

Table 1 Summary statistics for the latest version of the pig genome sequence (extracted from Ensembl database Oct 9 2012).

Assembly:	Sscrofa10.2, Aug 2011
Base Pairs:	3,024,658,701
Known genes	10,201
Novel genes	8,183
Predicted genes	3,256
Pseudogenes	380
RNA genes:	2,989
Gene exons	197,675
Gene transcripts	26,487
SNPs, indels	484,949

However, from an animal breeding and genetic improvement perspective, perhaps the most important number lies in the final line of the table. Already there are close to half a million sequence variants, with an average spacing of about 6.2kb, many of which have already been incorporated into large scale genotyping systems. And of course, many other sequence variants have been and continue to be discovered and documented in other studies. This means that at least some sequence variants are already known for all genes and future mapping of phenotypic effects to genomic and even gene regions is constrained only by the scope and cost of our genotyping systems.

Table 2. Summary statistics for chromosome 18, the smallest autosome (extracted from Ensembl database Oct 9 2012)

Length (bps)	61,220,071
Known Protein-coding Genes	269
Novel Protein-coding Genes	165
Pseudogenes	7
miRNA Genes	27
rRNA Genes	3
snRNA Genes	26
snoRNA Genes	7
Misc RNA Genes	6
Variations	12,868

Table 2 summarises the content of the smallest autosome which alone contains over 62 million nucleotides. As well as the protein coding genes, this summary categorises the important noncoding RNAs, many of which are now implicated in regulation of expression of other genes and the evolution of organismal complexity. MicroRNAs (miRNas) are important post-transcriptional regulators of gene activity and are involved in regulating the transition from stem cells to differentiated tissue. Small nuclear RNAs (snRNAs) associate with proteins to regulate gene expression, splice out introns and maintain telomeres. Small nucleolar RNAs (snoRNAs) guide the processing of other RNA molecules. And there are several other known categories of non-coding RNAs and probably many more to be discovered.

Progress in mapping economically important loci

27 years ago even the concept of a QTL did not exist. Genetic improvement programs in domestic animals depended on the application of quantitative genetics theory refined and developed from the original work of R.A. Fisher (1918). In this Fisherian theory, the genes underlying genetic variation in populations were hidden in a “black box”, unknown and unknowable. Nevertheless, based on the assumption of mass action of large numbers of genes of small effect, this theory permitted useful predictions of genetic merit and enabled very cost effective progress in selective improvement programs in many species.

In the pig, a couple of genes with mutations of large effect were known, most particularly at the so called *hal* locus, where a mutation of large effect predisposed to porcine stress syndrome with pleiotropic effects on meat quality. Not even the rudiments of a genetic map existed for this species and the genetic landscape of their 19 pairs of chromosome was *terra incognita.* A small number of genetic linkages were recognised but were of little practical utility.

However it gradually became clear that developments in genetic marker technology and improved genetic maps would enable the recognition and mapping of genetic regions with moderate sized effects on economically important traits. It was hoped that by lifting the lid on the black box, the identification of genes and chromosomal regions causally involved in variation would improve the efficiency of selection programs. The era of QTL mapping, mainly motivated by the objective of marker assisted selection (MAS) had commenced with pigs leading the way with the seminal paper by Andersson *et al.* (1994) reporting four QTL, the first ever in any domestic species. Figure 1 plots the rate of QTL discovery from 1994 through a peak of 1828 published in 2008, giving a current cumulative total of 7,451 for the pig.

However, despite this hugely impressive to genetic improvement programs has been hampered by the relative imprecision of QTL map positions and also the fact that many QTL were mapped in resource pedigrees generated by wide crosses between extreme populations, sometimes involving breeds of limited economic relevance to the commercial industry. The pursuit from QTL to QTN (quantitative trait nucleotide, namely the causal genetic mutation responsible for the phenotypic effect) continues, since genuine biological understanding lies in that direction. However, a fundamental shortcut to industry level exploitation of genetic markers arose from technical developments in large scale cost effective genotyping, allowing Meuwissen, Hayes and Goddard (2001) to firmly hammer the lid back onto the Fisherian black box. They demonstrated that with sufficient genetic markers genome-wide (in fact, many thousands), a training population to work out the relation between markers and phenotypes and the application of some sophisticated statistics, it was possible to effectively estimate breeding values based entirely on marker genotypes, without any prior knowledge of the marker effects or indeed even of their genomic position. Marker assisted selection (MAS) had mutated into genomic selection (GS), which exploited the linkage disequilibrium between markers and causal genes without any concern about what the causal genes were or where they were located. Thus mapping or even recognition of QTLs is not required for the implementation of genetic markers in genetic improvement programs. Substantial efficiencies in genetic improvement could follow. These included reductions in generation interval, since genomic breeding values could be estimated at birth and improvements in accuracy of selection for individuals with few progeny or other relatives.

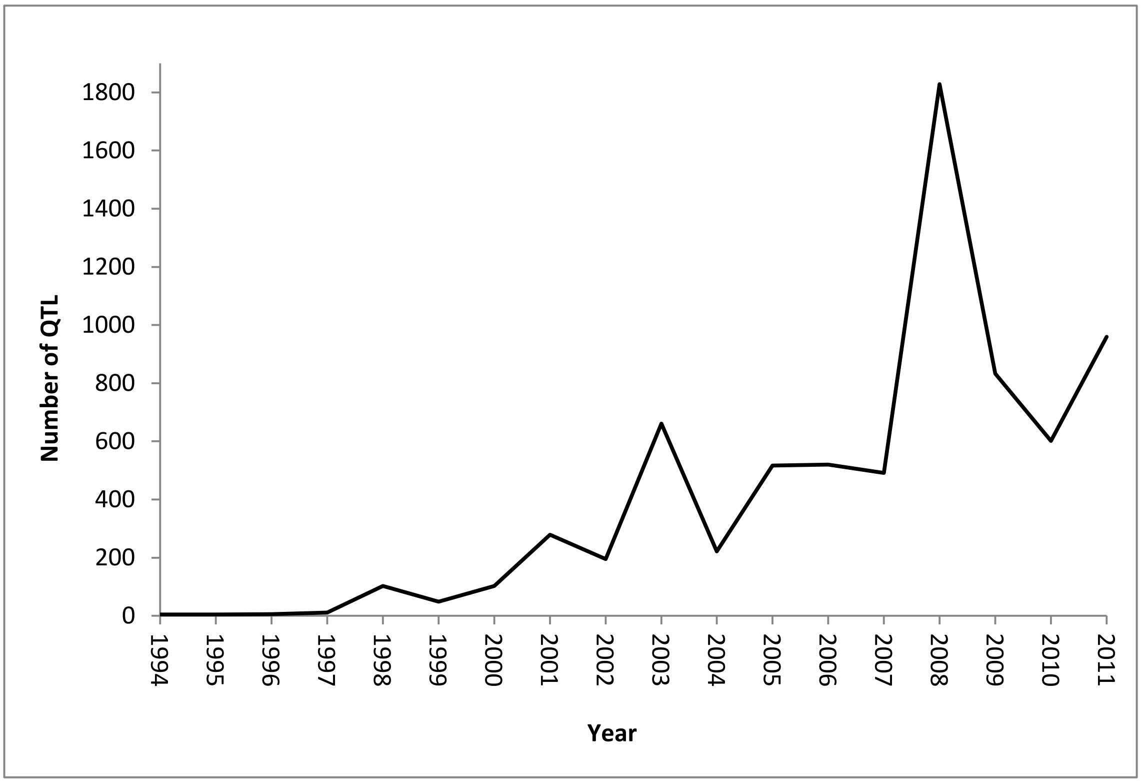

Figure 1. Rate of publication of pig QTLs (based on figures in Pig QTLdb in October 2012 – www.animalgenome.org/QTLdb)

Looking to the future – nextgen sequencing and beyond

There has been a dramatic reduction in the cost of DNA sequencing over the past 3 or 4 years. Traditional Sanger sequencing, which produced the human genome sequence in 2002 and the pig genome sequence in 2009, costs approximately $2400 per million bases of sequence produced. The currently most cost effective nextgen system, the IlluminaHiSeq2000, can now produce the same amount of sequence for $0.07 (Liu *et al*, 2012), a reduction of over 3400 fold, with the reasonable expectation of continued improvements towards a $1000 genome or even lower cost. Eventually, genome sequencing rather than the use of genotyping chips will become the favoured technique for genotyping. This will have the advantage of not requiring prior knowledge of SNPs for them to be detected and the high likelihood that the causal SNPs responsible for phenotypic effects on economically relevant traits will automatically fall out of the analysis. It won't necessarily be easy since the closer SNPs are to each other physically, the higher the level of linkage disequilibrium and many non-causal SNPs will be in perfect association with the few responsible for the effects. Already research aimed at exploitation of low coverage sequence and SNP imputation is examining the possibilities for genotyping by sequencing.

Remarkably, we are now on the cusp of a further revolution in sequencing based on reads from single template molecules. These new technologies have many advantages, most importantly long sequence reads capable of spanning repeat elements in the genome and more easily recognizing rearrangements of DNA such as inversions. Some of these technologies depend on pulling a single strand of DNA through a nanopore. As each nucleotide base passes through, electrical or optical properties can be recorded, which are distinct for each base. In fact, this technique can even recognize base modifications, such as methylation of cytosine, which are so important in epigenetic signaling. Many thousands of these nanopores can be read simultaneously. The Oxford Nanopore MiniIon (www.nanoporetech.com) is a device that can be attached to the USB port of a computer, does not require purified DNA and can generate almost a gigabase of DNA sequence.

A final vision for pig genetics

Genomic selection provides an excellent adjunct to conventional performance based selection in pigs. I envisage that with future major reductions in genome sequencing costs, very cost effective genome wide genotyping will be possible, enabling an optimal combination of conventional and genomic selection. However, the ability to use future-gen sequencing will also mean that eventually all underlying causal variants in genes will be exposed, even newly occurring ones. At that stage, we will have the combination of effective selection and the window into genuine biological understanding of the underlying processes. We will have taken the lid back off the black box. Of course constraints will still exist since the effects of some genes will still be too small for detection.

References

Andersson, L., C. S. Haley, H. Ellegren, S. A. Knott, M. Johansson, K. Andersson, L. Andersson-Eklund, I. Edfors-Lilja, M. Fredholm, I. Hansson, J. Håkansson and K. Lundström (1994) “Genetic Mapping of Quantitative Trait Loci for Growth and Fatness in Pigs.” Science 263: 1771-1774.

Archibald, A. L., C. S. Haley, J. F. Brown, S. Couperwhite, H. McQueen, D. Nicholson, W. Coppieters, A. Van de Weghe, A. Stratil, A. –K. Wintero, M. Fredholm, N. J. Larsen, V. H. Nielsen, D. Milan, N. Woloszyn, A. Robic, M. Dalens, J. Riquet, J. Gellin, J. –C. Caritez, G. Burgaud, L. Ollivier, J. –P. Bidanel, M. Vaiman, C. Renard, H. Geldermann, R. Davoli, D. Ruyter, J. M. Verstege, M. A. M. Groenen, W. Davies, B. Hoyheim, A. Keiserud, L. Andersson, H. Ellegren, M. Johansson, L. Marklund, J. R. Miller, D. V. Anderson Dear, E. Signer, A. J. Jeffreys, C. Moran, P. R. Le Tissier, Muladno, M. F. Rothschild, C. K. Tuggle, D. Vaske, J. Helm, H. –C. Liu, A. Rahman, T. –P. Yu, R. G. Larson and C. B. Schmitz (1995) “The PiGMaP consortium linkage map of the pig (Sus scrofa)”. Mammalian Genome 6: 157-175.

Fisher, R. A. (1918) "The Correlation between Relatives on the Supposition of Mendelian Inheritance." Philosophical Transactions of the Royal Society of Edinburgh 52: 399–433.

Groenen, M. A. M., L. B. Schook and A. L. Archibald (2011) “Pig Genomics” Chapter 8 in The Genetics of the Pig (Eds MF Rothschild and A Ruvinsky) pp179-199.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. (2012) “Comparison of next-generation sequencing systems.” Journal of Biomedicine and Biotechnology 2012: doi:10.1155/2012/251364

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard (2001) “Prediction of total genetic value using genome-wide dense marker maps.” Genetics 157: 1819-1829.