I HAVE A DRAFT GENOME FOR MY SPECIES…WHAT NOW?

Claire M. Wade

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

SUMMARY

As researchers, we need to question our objectives in gene mapping. Are we seeking merely a genetic test, or do we want to find the causative mutation so that we can understand the underlying biology. For a number of practical applications, a DNA test may be sufficient to assist the research sponsors, but as scientists, we have the opportunity to learn more about biology from our research. Additionally, it is becoming increasingly difficult to publish work that does not report a functional relationship with the association. This might be established through the discovery of an exonic mutation or through means such as expression analysis or other functional analysis if the mutations are regulatory.

Given that we have decided to proceed to mutation discovery, we require an effective approach. Our experiences with trait mapping in dog and horse have shown us that many mutations are regulatory, and that these can be elusive. Our use of target enrichment and sequencing with Massively Parallel techniques have shown us that there might be many mutations that are in LD with our trait of interest and that prioritizing these is the first step in establishing a functional basis to the phenotype.

INDRODUCTION

As we enthusiastically rush to gather the tools to enable genetic trait mapping in our species of interest that we often fail to stop and imagine what might be needed once we are successful in actually mapping something. In the past few years, more than 24 mammalian genomes have been taken to full draft coverage. The projects that generate the assemblies of these genomes frequently provide other essential resources such as SNP maps and the computational annotation of genes and features onto browsers that are easily accessible to the general public. These SNP resources have been used to produce genotyping arrays in most domestic species. This talk will focus on some lessons learned from array design and gene mapping thus far in the mouse, dog and horse. Some preliminary experiences with the use of Massively Parallel Sequencing (MPS) for targeted sequencing in the horse will also be discussed.

EFFECTS OF POPULATION HISTORY ON LINKAGE DISEQUILIBRIUM AND MAPPING SUCCESS

The architecture of linkage disequilibrium (LD) in mammalian genomes is a product of their particular population histories. In particular, factors leading to population bottlenecks have two large effects. First, population bottlenecks create founder effects that reduce the number of alleles that may occur at any given locus. Second, the bottlenecks have an effect of re-setting the “recombination clock”. That is, whole haploid chromosomes are forced into the new population and from that point onward, the alleles or haplotypes on those chromosomes are inherited as a unit until separated by recombination. The effect of this is to drastically lengthen LD in the short to medium term.

In the genome of the mouse, inbreeding and the creation of laboratory mouse strains from relatively few founders in the past 200 years has led to an extreme of this process. In essence, the inbred laboratory mouse strains are recombinant inbred lines of the different mouse sub-species, with the main contributors being *Mus musculus domesticus* and *Mus musculus musculus* but other contributors are *Mus musculus castaneus and Mus musculus molossinus*. This suggests that it should be relatively easy to map genetic traits in mice, but that the resolution of the mapping will be coarse. The LD across inbred laboratory mice measures over an average of two megabases (Mb). Even when a trait is successfully mapped, there is still much territory to sort through to try and find the mutation(s) responsible. Mutation detection through mapping in the mouse has not been extremely successful to date because of this problem of coarse resolution.

The population history of the dog provides advantages for gene mapping relative to the mouse. The history of dog domestication and breed creation is such that this species has been through two different major bottlenecks. The first bottleneck occurred at the point of human domestication of dogs. It is well established now that domestic dogs have been derived from Grey Wolves from Europe and Asia. Because wolves can be dangerous, relatively few individuals were taken from the wild to be tamed by humans. Those few that were taken forced a situation where relatively few chromosomes entered the dog population, but the long time period since domestication commenced (10-40K years ago) has recombined these few chromosomes very well so that if the domestic dog population is examined as a whole, the LD is very short. In humans the mean LD is regarded as short at 15-25 kilobases (Kb), while across all dog breeds it is even shorter and is of the order of 8-10Kb.

This would at first suggest that to undertake successful gene mapping in the dog, we will need more than twice as many genetic markers (most commonly used now are SNPs) than for mapping similar traits in humans. Fortunately, that is not the case. A second feature of dog genetic history is that during the Victorian era, humans became fascinated with the concept of dog breeding and showing. As a result, kennel clubs were formed and groups of dogs “breeds” that were genetically isolated from other domestic dogs were created. The effect of this second period of recent population bottlenecking is to create a situation in which the LD within a breed is of mouse-like proportions. In fact, the mean within-breed LD in the dog is of the order of one megabase. Indeed, this particular population history creates an ideal situation. If we have a mendelian trait that is segregating within a breed, then we can map the genes influencing the trait at coarse resolution with relatively few markers. Then if we wish to narrow the interval of association we can simply use other breeds that segregate the trait and make use of the short across-breed LD.

The genome of the horse has been recently completed. Horses are estimated to have been domesticated between 4,000 and 6,000 years ago and no populations of undomesticated horses exist in the wild other than perhaps the Przewalskii (Mongolian-wild) horse. This also has been shown to be introgressed with domestic stock. It seems that, unlike the dog, the domestication process of the horse has resulted in the capture of all the horses from the wild and so there is no true “domestication bottleneck”. The second bottleneck akin to the breed creation bottleneck in dog does however exist. But the genetic isolation of horse breeds is substantially reduced compared with dog breeds. An exception is the thoroughbred horse. The thoroughbred has been developed from few founders and has not allowed introgression from other breeds for a long period of time. The LD in this breed is much like that of a dog breed. The average within-breed LD in horses is approximately 150Kb and for the Thoroughbred is closer to 500Kb. Across breeds, there is much haplotype sharing and the LD remains relatively long at 50-75Kb. A SNP map of 1.2 million SNPs was generated as part of the horse genome assembly project. The SNPs were derived from light whole genome shotgun sequencing of 9 horse breeds sampled from the horse populations of Europe, the Americas and Asia.

THE RIGHT TOOL FOR THE JOB: GENOTYPING ARRAYS

All of the aforementioned species have genotyping arrays now available. For the dog, arrays were designed on both Affymetrix and Illumina iSelect platforms. The long LD within dog breeds suggested that between 15 and 20,000 SNP would be needed for trait mapping. Affymetrix arrays holding initially 63K and later 128K SNPs were designed and an Illumina iSelect array of 23K SNPs that had been validated on the Affymetrix arrays.

The Affymetrix platform uses a restriction digest of the genomic DNA with either one or two restriction enzymes (usually only one for non-human species). The fragments from the restriction are amplified and sized and a particular size range of fragments is used to take part in the genotyping. This strategy has two disadvantages. First, in the design phase of the array, the number of SNPs occurring in parts of the genome residing in these expected fragments is much reduced. Only about 10% of the discovery SNPs can be used for the assay design. On this system, the most efficient approach is to pre-digest the genome with the chosen enzyme and then to sequence only the fragments in the desired size range for SNP discovery at the beginning. Second, the restriction process requires that the DNA quality should be very good and this typically results in an unsuitability of samples derived from buccal swabs or hair samples for whole genome genotyping purposes. This is because DNA degradation pre-cuts the genome, so that when the restriction digest is carried out the DNA in the desired fragment size range may not be the DNA that you are expecting and array performance is extremely poor. The high genotyping success rates reported for arrays such as the Human arrays are never observed on non-human species arrays because the SNPs used for non-human mammals are not usually pre-validated on the Affymetrix system. For the mouse arrays, two 256K SNP arrays (512K SNP total) yield 148K usable, polymorphic SNP. For the dog arrays, the 63K SNP array yields 27K usable, and the 128K array design yields 50K.

The Illumina iSelect platform makes use of whole genomic DNA and so is relatively unaffected by sample degradation. Also, because it does not require pre-digestion of the genome in the array design, all known SNPs are available for array design. The yield from the arrays is very good – typically of the order of 85-90%. For the dog Illumina array, 24K SNP were designed for a yield of 22K SNP. For the horse array, exactly 60K SNPs were designed and the yield is 54K. This array is considerably more expensive than the Affymetrix platform and the array processing facilities are less available but the data quality is exceptionally good.

In 2007 the horse research community formed a consortium to produce a horse genotyping array. Power calculations suggested that an array of 150,000 SNP was desirable. While the Affymetrix platform offered 1 million SNP designs in an affordable package, the horse community only had 1.2 million polymorphisms available for design of which only 10% were expected to co-occur with the fragment sizes produced by restriction digest. Thus the only Affymetrix option available was the 128K design with an expected yield of 50K SNPs. The Illumina 60K option had a similar expected yield. The Illumina iSelect genotyping system was chosen for its high data fidelity and for the capacity to use the greatest proportion of the SNPs available. The expense of this array meant that the community was unable to afford more than 60,000 bead-types but the samples used by the community came from predominantly swab or hair DNA and so were subject to degradation.

SUCCESS MAPPING WITH GENOTYPING ARRAYS IN THE DOG

Successful coarse-resolution mendelian trait mapping in the dog has been carried out with as few as 7 cases and 10 control animals. Because of the structure of the genome in dog populations, there is the advantage that if the same trait segregates in multiple breeds, the other breeds can be brought in for fine mapping to reduce the interval of association. Of course, the density of markers required for fine mapping is very much greater than that of the genome-wide mapping runs. There are several examples of the successful use of this approach in the dog (Karlsson *et al.* 2007; Salmon *et al.* 2007; Drogemuller *et al.* 2008; Awano *et al.* 2009).

A different breed segregating the trait of interest is, however, not always available and this can create a substantive impediment to mutation detection. Even so, SNP and other polymorphisms within the broader interval can be used for robust genetic testing if mutation discovery can be delayed as is the case for the test for Hyperparathyroid tumors in Keeshonden (http://www.akcchf.org/news/index.cfm?article_id=145). Otherwise, alternate bio-informatic approaches can be used to identify candidate genes in the interval of association as has been used for mutations associated with Degenerative Myelopathy and Rod-Cone dystrophy in dogs (Wiik *et al.* 2008; Awano *et al.* 2009). In both cases, literature searches were used to identify candidate genes within the interval of interest and exonic mutations were identified. In the case of the Keeshonden, the causative mutation remains elusive although the genetic test is 100% effective in this breed. Of the traits described, the mutations of three are exonic (Rod-Cone Dystrophy and Degenerative Myelopathy, Chinnese Crested), two are regulatory (White boxer, ridge) and one is unknown (Keeshonden). We expect that many non-lethal mutations will be regulatory in nature.

PROOF OF PRINCIPLE MUTATION DETECTION IN THE HORSE INTEGRATING MASSIVELY PARALLEL SEQUENCING

As part of the horse genome analysis (Wade *et al.* 2009) we chose to integrate positional mapping with new sequencing technologies including MPS to gain an idea of the likely success of these technologies for mutation detection. To do this we chose to study four mendelian coat-colour traits, but only three were successful because two horses with the fourth phenotype failed in sequencing. In each case, exons within the mapped interval had been already been assayed by limited sequencing with PCR and no exonic mutations had been identified. For this reason the mutations were expected to be regulatory.

While MPS is a very cheap way of generating sequence, it is indiscriminant with respect to target. Eight lanes of sequence with the Illumina Genome Analyzer generates about 2Gb of random sequence and costs about $15K US. If you give the sequencer genomic DNA, then you will get light cover of the entire genome and perhaps considerable cover of mitochondrial DNA depending on the tissue from which DNA was extracted. The sequences produced by MPS have considerable error rates with error types that vary depending on the platform used. This means that you must have at lease five fold cover of the target to reliably call SNP or Insertion-deletion events and possibly even more cover to call copy number variants. A more efficient method of carrying out mutation discovery using MPS is to enrich the sequencing for target DNA so that the cover can be effectively increased and more individuals (including replicated affected and unaffected individuals) can be sequenced.

Achieving effective enrichment of sequence in the target region presents one of the major challenges of the new sequencing technologies. A number of methods of target enrichment are available. Some involve long-range PCR and others involve target capture using tiled genomic sequence from a draft genome. Long-range PCR is suitable if it is expected that there will be considerable divergence from the draft genome, or when no draft genome is available. For the horse analysis we used a hybrid capture technique. For this, we sent draft-genome sequence from the target region to Nimblegen and had high resolution comparative genome hybridization slides constructed for the four regions. All four regions were on the same hybridizing array. DNA from affected horses for the four traits was hybridized to the slides and the slides were washed to remove the extraneous DNA. Next, the hybridized DNA was eluted from the slide and amplified using PCR. The amplified products were sent for Illumina Genome Analyzer II (Solexa) sequencing.

In horse, the strategy employed was to take regions that had been previously identified to harbour genes influencing the traits by genetic mapping with microsatellite markers. The regions to be assayed varied considerably in size, with the smallest being 300Kb and the largest 10Mb. We used Sequenom Mass Spectrometry genotyping with one or two pools of SNPs (each pool may contain up to 35 SNP) and many horses from different breeds to reduce the intervals to manageable size for MPS. To do this effectively we needed sufficient density of SNP to see haplotype breaks in individual horses. In the horse, this ideally requires at least 5 SNP per 100kb. For the largest region, to save funds we gambled and focused the fine-mapping in the vicinity of a gene that had an expression difference and so we used sparser than desired genotyping over 2MB. If this had not revealed an association we would have broadened the search but this was not required. In each case, by fine mapping we were able to reduce the interval of association to 200-300Kb. These intervals were tiled on hybridizing arrays and became our sequencing targets.

The success of this approach was varied. The best performing samples achieved coverage of 150× of the target region with 70% of all sequence tags falling within the enriched region. This was from 1 lane of sequencing (tag size 35 base pairs). The worst sample had 0.2× cover of the target (mainly in repetitive sequence). Alignment to draft genomic sequence was carried out by three methods. MAQ (Li *et al.* 2008), Spines-aligns (Maucelli pers. comm.) and Smatch (Kirby pers. comm.). The latter two methods are under development at the Broad Institute of Harvard and MIT. All of these alignment methods worked well over the limited regions assessed. MAQ is freely available on the web. The samples with low cover were determined to be primarily affected by hybridization failure.

One of the traits assessed (Grey) was known to be associated with sequence duplication and this was readily detected in the normalized sequence tags by assessing mean draft genome coverage base by base. As expected, many mutations (predominantly SNP and some insertion-deletion events) were observed to be in LD with the expected associations. To prioritize these mutations for functional study, we made use of transcriptome profiling data(Coleman 2009) and also conserved element analysis (Garber *et al.* 2009). Briefly, the transcriptome profiling resulted from mRNA-seq using Illumina Genome Analyzer II on eight horse tissues (none were skin). The conserved elements analysis is the result of multiple alignments of sequences from 24 mammals with at least draft genome (7×) coverage. The conserved elements are detected with word sizes of 8 bases or more. These are considered to be regulatory elements of evolutionary significance. At the time of writing, two of these mutations are maintaining association over a larger set of horses from the affected breeds.

I HAVE A DRAFT GENOME FOR MY SPECIES…WHAT NOW?

For many genomes so far, there has been sufficient community fundraising to enable the production of large scale commercial genotyping arrays. Typically the funding required to create an array is around $1million USD and so this may be beyond the reach of species without commercial or public importance. Once it is decided to create an array, the choice of array technology must be influenced first by the population genetics of the species. For species without recent population bottlenecks, it is likely that a large number of SNP will be required to map even mendelian traits effectively, and should sufficient SNP exist, care must be taken to identify those that can be successfully used on the genotyping platform chosen and the tissue types that will be commonly used for DNA extraction. For communities unable to generate sufficient funding for commercial array production other methodologies are required, such as large scale hybridization arrays combined with MPS technologies. At this time the success of such approaches can only be speculated but with improvements in assembly of paired-end MPS data e.g. Velvet (Zerbino and Birney 2008) a draft genome may not even be required.

Given sufficient SNP density, mapping projects using commercial large scale genotyping arrays will inevitably lead to the successful mapping of mendelian traits at the very least. Depending on the distance to the last strong population bottleneck, the resolution of this mapping might be quite coarse and the researcher may be left with a considerable region to analyse to discover the causative mutation. This region should ideally cover unassociated flanking sequences to ensure that the mutation lies within the assayed interval. Affected and unaffected individuals should be sequenced as replicates with sufficient coverage to enable confident mutation detection. Fine mapping can be used in multi-breed data to effectively reduce the size of assayed region so that coverage may be increased.

As MPS becomes cheaper it may be feasible to sequence an entire region of association from within a single breed affordably or possibly the entire genome. Given that in our limited regions we typically identified more than 100 associated mutations, the number of potential mutations that would be identified by larger scale sequencing would be truly staggering and require significant bioinformatic analysis for everything from assembly to alignment and the detection of high quality disparities between affected and normal individuals. Once disparities are identified, rational approaches must be employed to reduce the search space for the causative mutation. Bioinformatic approaches based on literature review, transcriptomics and conservation have the ability to reduce the search space by more than 90%. Enormous amounts of data result from even limited use of MPS technologies. Tera-/Peta-byte levels of space for processing and storage of data are required along with the computational expertise and equipment to analyse the data. These represent significant impediments to the successful application of MPS without target resolution in the short to medium term.

REFERENCES

Awano, T., Johnson T.G., Wade, C.M., Katz, M.L., Johnson, G.C., Taylor, J.F., Perloski, M., Biagi, T., Baranowska, I., Long, S., March, P.A., Olby, N.J., Shelton, G.D., Khan, S., O'Brien, D.P., Lindblad-Toh K. and. Coates J.R. (2009) Proc Natl Acad Sci U S A **106**: 2794

Coleman, S. J., Zeng S., Wang K., Khrebtukova I., Luo S., Mienaltowski M.J., Schroth G., Liu J., and MacLeod J.N. (2009) I*n preparation.*

Drogemuller, C., Karlsson, E. K., Hytonen, M. K., Perloski, M., Dolf, G., Sainio, K., Lohi, H., Lindblad-Toh, K., and Leeb, T. (2008). **321**: 1462.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman N. and X. Xie (2009) *Bioinformatics* **25**: i54

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H., Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R., Kulbokas, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler, H., Kampe O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson, L. and Lindblad-Toh, K. (2007). *Nat Genet* **39**: 1321

Li, H., Ruan J. and Durbin, R. (2008) *Genome Res* **18**:1851

Salmon Hillbertz, N. H., Isaksson, M., Karlsson, E. K., Hellmen, E., Pielberg, G. R., Savolainen, P., Wade, C. M., von Euler, H., Gustafson, U., Hedhammar, A., Nilsson, M., Lindblad-Toh, K., Andersson L. and Andersson, G. (2007) *Nat Genet* **39**:1318

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R. R., Blöcker, H., Distl, O., Edgar, R. C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J. N., Penedo, M. C. T., Raison, J. M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D. F., Biagi, T., Binns, M. M., Chowdhary, B. P., Coleman, S. J., Della Valle, G., Fryc, S., Guérin, G., Hasegawa, T., Hill, E. W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J. R., Murray, J., Nergadze, S. G., Onofrio, R., Pedroni, S., Piras, M. F., Raudsepp, T., Rocchi, M., Røed, K. H., Ryder, O. A., Searle, S., Skow, L., Swinburne, J. E., Syvänen, A. C., Tozaki, T., Valberg, S. J., Vaudin, M., White, J. R., Zody, M. C.,Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team, Lander, E. S., and Lindblad-Toh, K. (2009) *(submitted)*

Wiik, A. C., Wade, C., Biagi, T., Ropstad, E. O., Bjerkas, E., Lindblad-Toh, K. and Lingaas, F. (2008) *Genome Res* **18**:1415

Zerbino, D. R. and Birney, E. (2008) *Genome Res* **18**:821