Canine Genetic Testing – Introduction to the Technologies.

Matthew Breen PhD CBiol FRSB,
Oscar J. Fletcher Distinguished Professor of Comparative Oncology Genetics
North Carolina State University

Pamela S. Rosman
AKC Delegate, Canaan Dog Club of America

Mark Dunn
Executive Vice President, American Kennel Club

This is a brief introduction to current technologies as they apply to canine genetic testing. The reader is directed to three short articles written by Dr. Breen that serve as a basic introduction to canine genetics, available at https://www.akc.org/author/matthew-breen/.

As canine genetic testing is becoming more widely available, whether via a veterinary health professional or direct to consumers, it is important for all dog fanciers to be familiar with the technologies used to deliver such testing. All dogs share the same collection of approximately 20,000 genes organized into the characteristic set of 39 pairs of chromosomes in each cell. However, at the DNA sequence level there is considerable variation (polymorphism) between individuals. It is this sequence variation that forms the basis of genetic testing for parentage verification, risk/detection of disease, breed ancestry, and forensic analysis. All genetic testing aims to match genetic signatures of an individual associated with one or more traits or phenotypes.

- **Biological sampling and DNA isolation.**

All DNA-based canine genetic testing begins with the collection of a biological specimen from the dog being tested. The type of biological sample, together with the manner in which it is obtained, stored, transported, and ultimately processed by a laboratory, are all key factors that can impact the quality and quantity of the resulting genetic data. For canine genetic testing the most common sources of biological material are blood and oral swabs of the mouth (buccal/cheek swabs). While buccal swabs are relatively easy to obtain, the quality and quantity of canine genomic DNA recovered is highly variable, depending on the type of swab used, the effectiveness of the buccal collection, and the cleanliness of the dog's mouth. In contrast, genomic DNA obtained from a whole blood sample is consistently of higher quality and quantity than from buccal swabs. However, obtaining a blood sample generally requires a licensed health professional, which can add time and cost. Modern approaches to DNA isolation from either blood or buccal swabs make use of barcoding and semi-automated processing to maximize efficiency and accurate sample tracking.

- **Genetic testing methodologies.**

Once genomic DNA has been isolated from a dog, it is analyzed by one or more techniques depending on the type of genetic variation being assessed. In 2020, the most common variants used in canine

genetic testing are short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs). Both types of genetic variation are inherited and passed from parents to offspring.

Short Tandem Repeats (STRs): A DNA sequence comprises four nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C), and it is the specific order of these nucleotides that determines the genetic code. A short tandem repeat (STR) is a sequence of DNA, generally 2-5 nucleotides in length, that is repeated numerous times at a specific location (locus) in the genome. STRs are named by the number of nucleotides in the repeat sequence, so 2, 3, 4, and 5 nucleotide repeats are referred to as di-, tri-, tetra-, and penta-nucleotide repeats, respectively. Each individual has two copies (alleles) of each STR, one copy inherited from each parent. The identity of each allele is referred to as the genotype. Where the two alleles of an STR inherited from each parent are the same, the genotype is referred to as homozygous. If the two alleles are different, the STR genotype is referred to as heterozygous. The polymorphism of an STR is due to the presence of a variable number of copies of the repeat element that occurs in different individuals.

Traditionally, the genotype of STR alleles is determined by analysis of a small section of the genome (less than 500 nucleotides) surrounding each STR. The technology used is referred to as polymerase chain reaction, or PCR, which is a process that makes millions of identical copies of a specific region of the genome. This process is called amplification and the region of the genome that is copied is referred to as a PCR amplicon. The genotype of each STR is reported as the size of both amplicons. For an individual STR there may be 5, 6, 7, or more alleles in a population, and so the possible number of combinations of the two alleles in any one dog can be high. The frequency of each allele in a population is used to calculate the degree of polymorphism of the STR in that population. Generally, the greater the polymorphism, the more useful the STR is for parentage testing. To maximize the ability for parentage exclusions, panels of STRs are combined. The current AKC parentage panel (SuperPlex-G) comprises 13 STRs as well as a sex determination marker, and provides over 99% confidence for parentage exclusions. However, STRs have limited utility in genetic health testing.

Single Nucleotide Polymorphism (SNP): Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in the canine genome, estimated to number several million per individual dog and tens of millions across breeds. A SNP represents an alternate nucleotide at a specific location in the genome, for example a ‘C’ (cytosine) in one individual may be a ‘T’ (tyrosine) in another. In contrast to the high degree of polymorphism of STRs, SNPs usually have just two alleles and so each has limited polymorphism. When used for parentage testing, many more SNPs (panels of more than 100) need to be assessed to provide the power of exclusion offered by fewer STRs (panels of fewer than 20). However, with millions of SNPs to choose from, combined with technology that can rapidly genotype thousands of SNPs in parallel, SNPs have become the variant of choice for parentage testing in many animal species. While most SNPs are benign, where a SNP does occur in a key sequence that determines gene function, there can be a health impact, such as higher risk of a particular disease. In addition to accurate parentage testing, SNPs can thus inform the inheritance of disease risk and aid in health management of individual dogs and breeding programs. The large number of SNPs distributed throughout each canine genome offers a valuable resource in facilitating ongoing research studies to

identify new segments of the genome associated with specific traits, including disease. Panels of trait-associated SNPs form the basis for most of the genetic tests available to dog owners in 2020.

There are several molecular technologies that can be used to obtain canine genetic test results, including PCR, array based technology, and direct sequencing of DNA. The choice of technology used determines how many SNPs can be analyzed for each canine DNA sample and also the number of samples that can be processed in a single batch (sample throughput).

PCR: For genetic tests performed to evaluate short DNA sequences, or small groups of sequences, PCR is commonly used as a reliable means to selectively amplify small segments of genomic DNA. For example, while the full genome of a dog is 2.4 billion nucleotides in length, the length of a region needed to determine the genotype of a single locus of interest may require analysis of a fragment of DNA sequence less than 100 nucleotides in length (i.e. 0.000004% of the genome). Analysis of the nucleotide content of each amplicon provides the genotype of the SNP being investigated. This approach is ideal for analysis of an individual SNP, but is not suited for genotyping the large numbers of SNPs required for genetic health testing.

DNA microarrays: Collections of known genetic variants may be consolidated as short reference DNA segments (50-100 nucleotides) immobilized on beads or on silicon/glass surfaces. These collections, referred to as genomic microarrays, provide a reliable and highly cost-effective technology to determine genotypes of large numbers of SNPs simultaneously. When the DNA of a dog is applied to a genomic microarray, the technology deciphers the exact genotype of each allele for multiple SNPs. Generally, microarrays used for genetic testing comprise a collection of well-defined variants that contribute to and/or cause disease, supplemented with a much larger set of SNPs evenly spaced throughout the whole genome. The number of variants represented on an array is limited only by the platform used.

Genotyping by microarrays offers an ideal technology for assessing large numbers of SNPs in high numbers of dogs at the same time (high throughput). The content of a genotyping microarray is determined after research has been performed to identify which SNPs should be included. A major advantage is that since the content is known, each SNP evaluated is the same for all DNA samples being tested. A limitation is that genotyping microarrays can only report on the content offered and so any new variant(s) discovered require development of a new microarray.

Genotyping by sequencing (GBS): Genotyping by sequencing (GBS), or targeted genotyping by sequencing, is a method used to identify the genotype of numerous SNPs at the same time. When used for genetic testing, the approach uses a DNA sample with much lower complexity than a whole genome sequence. Lowering the complexity of the DNA is achieved through a process of selective enrichment of short DNA sequences flanking the SNPs of interest. The enriched segments of the genome are then subject to simultaneous DNA sequencing to reveal the genotype of each SNP in the panel. Identification of the genotypes requires the DNA sequences to be analyzed using sophisticated computational tools designed to optimally acquire, store, and process the data. Spanning the interface of biology and computational science, these tools are termed bioinformatics.

Low pass whole genome sequencing (low pass WGS): The final cost of generating a high-quality sequence of the human genome was estimated to be almost $3 billion. A major component of the human genome project was a huge investment in new technologies that led to the cost of DNA sequencing declining substantially each year. In 2020, while some companies claim to be able to generate whole genome sequences for under $100, these are not full genomes with every nucleotide sequenced and accurately identified, rather they are partial genomes. High-coverage whole genome sequencing (WGS), where every nucleotide is sequenced independently at least 15-25 times (referred to as 15x-25x WGS), is still the gold standard approach. However, the cost of this approach remains in excess of $1,000 per genome and so is prohibitively costly for routine adoption. To achieve a more affordable alternative, low pass whole genome sequencing (<1x WGS) is now being used, with a typical cost of less than $200 per DNA sample. In low pass WGS, not all nucleotides comprising the genome are sequenced even a single time (hence the <1x) and so the resulting sequence is littered with missing information and potential errors. A computational process, referred to as imputation, is used to fill in the gaps with the most likely nucleotides and also reduce errors. Imputation is a form of analysis where the genotypes of SNPs tested (observed data) are used to estimate the genotypes of neighboring SNPs that are not obtained directly (unobserved data). This process relies on knowledge of how often the genotype of the tested SNP is observed with the genotype of the neighboring SNPs in previously generated genome data. While imputation can be very powerful, its accuracy is dependent on having sufficient knowledge of the genome of the species being tested.

In addition to low pass WGS providing millions of SNP genotypes for a similar cost to a genotyping array, this approach requires much less DNA than needed for a microarray. As with all emerging technologies, evaluating all aspects of performance are important to determine feasibility and, in this context, SNP genotyping accuracy. As low pass WGS becomes more widely adopted for canine applications, the lower costs, reduced DNA requirement, and improved accuracy through more mature bioinformatics tools, will promote this approach for routine canine genetic testing.

SUMMARY

Genotyping is a powerful tool when applied to canine genetic testing. Traditional methods of genotyping for parentage testing have used STRs, which are analyzed on the basis of the size of each of the two alleles. Consideration of the genotypes of multiple STRs increases the combined power of canine parentage exclusion to an accuracy of over 99%. While this is a very reliable approach, it is a dated technology that remains costly with limited sample throughput. A shift towards a less costly method to determine genotypes is desirable, especially if the technology also allows many more samples to be analyzed simultaneously. The use of a SNP based genotyping system offers key advantages over existing methods. SNPs allow for more rapid and reproducible genotyping options compared to the use of STRs, with current platforms allowing the side by side evaluation of substantially more variants. Unlike STRs, which are typically neutral, SNPs may be associated with specific inherited canine traits, including disease. Although the current AKC DNA program is focused on parentage testing, it is worth considering new opportunities to leverage these technological advances. Whether by an array based or a

sequence based approach, the use of SNP technology to offer combined canine parentage and health testing is now a feasible option to consider.