8-2020

Next-Generation Sequencing in Equine Genomics

Jessica L. Petersen
University of Nebraska-Lincoln, jessica.petersen@unl.edu

Stephen J. Coleman
Colorado State University, stephen.coleman@colostate.edu

Published in *Veterinary Clinics of North America: Equine Practice* 36:2 (August 2020), pp. 195–209;
doi: 10.1016/j.cveq.2020.03.002

Published online July 9, 2020.

Next-Generation Sequencing in Equine Genomics

Jessica L. Petersen[1] and Stephen J. Coleman[2]

1. Department of Animal Science, University of Nebraska–Lincoln, Lincoln, Nebraska, USA
2. Department of Animal Sciences, Colorado State University, Fort Collins, Colorado, USA

Corresponding author – Stephen J. Coleman, Department of Animal Sciences, Colorado State University, 1171 Campus Delivery, Fort Collins, CO 80523-1171, USA, email stephen.coleman@colostate.edu

Abstract

- Next-generation sequencing of both DNA and RNA represents a second revolution in equine genetics following publication of the equine genome sequence.
- Technological advancements have resulted in a wide selection of next-generation sequencing platforms capable of completing small targeted experiments or resequencing complete genomes.
- DNA and RNA sequencing have applications in clinical and research environments.
- Standards for the validation and sharing of next-generation sequencing data are critical for the widespread application of the technology and applications discussed herein.
- As researchers and clinicians develop a better understanding of how genetic variation and phenotypic variation are linked, next-generation sequencing could help pave the way to personalized and precision management of horses.

Keywords: genomics, equine, RNA sequencing, transcriptome, genetic variation

Introduction

The sequencing and assembly of a reference genome for the horse has been revolutionary for investigation of horse health and performance. Since its publication,[1] the reference genome has enhanced and accelerated genetic research in the horse, led to the development of new ideas regarding management and precision medicine, and has led to the development of

powerful tools that increased the scope and resolution of understanding the genetic underpinnings of equine physiology and disease pathology.[2,3] The insights gained into equine health as a result of these new tools and ideas are expertly reviewed in the accompanying articles of this special issue. The advent and application of next-generation sequencing (NGS) methods represent a second revolution for the study of equine genetics, enabling researchers to exploit and explore the information encoded in the equine genome through their experiments. NGS has also improved the ability of researchers to translate their discoveries into clinically relevant applications. This article provides an overview of the history and development of NGS, details some of the available sequencing platforms, and describes currently available applications in the context of both discovery and clinical settings (Fig. 1).

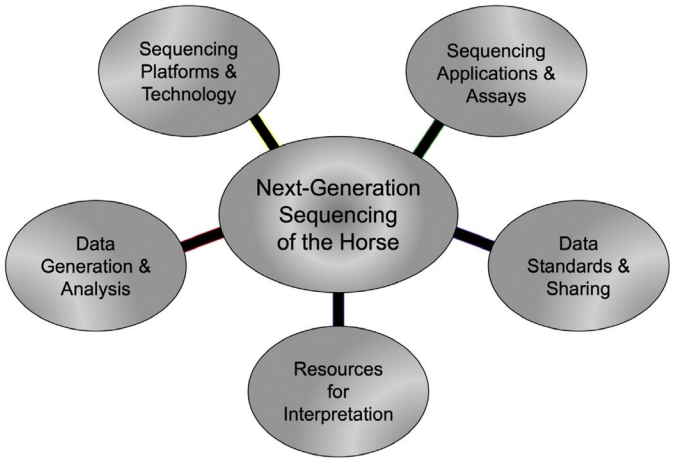

Figure 1. Visual summary of the key points for NGS application in the horse in both discovery and clinical settings.

Building Genomic Resources for the Horse

The use of DNA sequencing to investigate the underlying cause of heritable conditions in the horse dates to the early 1990s. At that time, before the development of an equine reference genome, genetic studies relied on the use of genomic information from other species to inform the investigation for important traits of the horse. Major successes using that approach include the identification of a missense mutation causative of hyperkalemic periodic paralysis in the quarter horse[4] and lethal white overo syndrome in American Paint Horses.[5] In 1995, the scientific communities' focus on generating genomic tools specific to the horse incited the formation of the Horse Genome Project. Through this collaboration, intentional and international partnerships were built across academic and industry institutions, resulting in the generation of comparative, linkage, and radiation hybrid maps of the equine genome (reviewed in Chowdhary[6]).

The most notable advancement for equine genomics thus far dates to 2006 when the National Human Genome Research Institute of the National Institutes of Health identified

the horse as a species of priority for genome sequence assembly efforts. In 2007, a draft reference equine genome was completed.[1] This reference genome, named EquCab2, was generated with sequencing data from a single thoroughbred mare, Twilight, resulting in an assembly with approximately 6.8-fold coverage. The assembly of these data was complemented by additional sequence information (bacterial artificial chromosome sequencing) of Twilight's half-brother, Bravo. At the time, the accuracy of Sanger sequencing and availability of linkage and physical maps of the genome resulted in EquCab2 being one of the highest-quality reference genomes of any agricultural species. The genome was estimated to be 2.7 billion base pairs (bp), with more than 20,000 protein-coding genes annotated in the initial effort.[1] This resource served as the basis for the development of genomic tools and discovery for the following decade with assays to detect genomic and transcriptomic variation in the horse[2,3,7–9] anchored in EquCab2.

In 2018, a new reference assembly, still based primarily on the sequence of Twilight, was released.[10] This improved reference genome, EquCab3, was the product of new technologies for sequencing of longer reads, helping to characterize repetitive regions of the genome. The EquCab3 assembly also incorporates data generated by methods that use structural proximity of sequences to help build continuity (Chicago[11] and HiC[12] libraries). Compared with EquCab2, EquCab3 has 90% fewer gaps, better coverage of GC-rich regions, which often include gene promoters; and more complete coverage of the transcriptome.[10] EquCab3 now serves at the primary reference genome assembly for the horse and should be used for the analysis of future sequence data. It continues to be improved through additional efforts to annotate not only protein-coding regions, but noncoding RNA as well as regulatory features.[13] Both EquCab2 and EquCab3 are available through the National Center for Biotechnology Information (NCBI), Ensembl, and University of California Santa Cruz genome browser utilities.

Technology

The driving force behind many of the developments in equine genetics, including but not limited to the reference genome sequence, has been ever-improving and increasingly accessible DNA sequencing technology. The advancing technologies are generally grouped into distinct generations by the scientific community to recognize the transformational impact they have had on the understanding of genetics. The history and impact of each generation of sequencing technology have been reviewed in detail.[14,15] This article presents a brief overview of each sequencing generation and specifically how it has or can affect studies of the equine genome related to animal health.

The method used to generate the data for assembly of EquCab2, data that were also used for EquCab3, was Sanger sequencing, first published in 1977.[16] Still used for projects concerning a single gene or small portion of DNA, Sanger sequencing produces a high-quality sequence in long fragments. Sanger sequencing relies on the selective incorporation of dideoxy nucleotides during elongation of the nascent DNA strand during in vitro DNA replication. Fragments are then visualized using an electrophoretic system to identify each nucleotide in sequence. This method, which can generate sequence fragments of about 800

bp, may be limited in throughput but remains the gold standard for accuracy (reviewed in Shendure and Ji[17]).

As researchers began to work to develop reference genome assemblies, increasing the throughput of sequencing technologies became a priority. The motivation behind the rapidly evolving technology was to improve access by increasing accuracy and data-generating capacity while at the same time decreasing costs. NGS technologies were designed to increase the rate by which data were generated through platforms that allowed for multiple sequence reads to be collected at 1 time and also by coupling the inclusion of labeled nucleotides with the step of reading their identity. Next-generation, or massively parallel, sequencing, therefore, had an advantage in its ability to generate a significantly greater amount of sequence data at 1 time, although read length was compromised compared with that possible with Sanger sequencing. In the past 10 to 15 years, NGS has become a standard method used in questions regarding the evolution of the species, for discovery of variation associated with phenotypes of interest, for the identification of diversity among individuals and breeds, and for identification of genome function associated with disease.

Sequencing Platforms

As sequencing technologies have advanced, sequencing platforms available to generate data have expanded at an astounding rate. Ten years ago, there were only a few types of instruments available; these were expensive and required significant laboratory resources to deploy. At present, there is a wide selection of instruments tailored to generate anywhere from a small amount of targeted sequence data to massive amounts of sequence data capable of characterizing an entire genome in a single experiment. The various platforms use different types of chemistry; those differences have been previously reviewed.[18,19] Now there is an NGS platform for most any job. As this technology continues to become more accessible and manageable, it enhances the opportunities for sequencing and its many applications to find their way into clinical practice.

The available NGS platforms can be classified into 3 main groups: production, benchtop, and portable systems. The choice of which platform to use depends on several variables of interest, the overall throughput needed, the accuracy of base calls, read length, speed of data generation, and budget. Each platform category is described briefly in the following paragraphs and is summarized in Table 1.

Table 1. Comparison of production, benchtop, and portable next-generation sequencing platforms

	System Scale					
	Production		Benchtop		Portable	
Descriptors	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Sequence output	7.5 Gbp	6 Tbp	1.2 Gbp	150 Gbp	1.8 Gbp	30 Gbp
Read output	0.5 M	20 B	4 M	400 M	7 M	12 M
Read length	50 bp	1 KB+	50 bp	1 KB+	—	10 Kb+
Platforms	Illumina HiSeq 4000 Illumina HiSeqX Illumina NovaSeq 6000 PacBio Sequel/Sequel II Oxford Nanopore PromethION		Illumina MiSeq Illumina NextSeq Illumina iSeq 100 ThermoFisher Ion S5/S5 XL Oxford Nanopore GridION		Oxford Nanopore MinION Oxford Nanopore Flongle Oxford Nanopore SmidgION	
Applications	Whole-genome sequencing Exome sequencing Targeted sequencing Epigenetic sequencing DNA-protein interactions Transcriptome sequencing Gene expression profiling Small RNA sequencing		Targeted sequencing Targeted expression profiling Small genome sequencing Small RNA sequencing		Small genome sequencing Targeted sequencing Targeted expression profiling Epigenetic sequencing	

Production systems represent the highest-throughput technology available and are targeted primarily for discovery and research applications. The designation of a production system is derived from the idea that a researcher would need to produce a genome or transcriptome sequence. Systems in this category have sufficient capacity to sequence an entire genome in a single run (realistically many genomes given the coverage needed). They are almost exclusively housed in core or service facilities because of the cost to purchase, deploy, and operate them. The advantage of these platforms is the output. One of the highest-throughput systems currently available, the Illumina NovaSeq 6000, can generate 6 Tb of sequence data or 20 billion reads in less than 2 days (https://www.illumina.com). This amount of data represents enough sequence to characterize the genome of 1 horse more than 300 times. More practically, this amount of data can be used to sequence the genomes of 15 individual horses to coverage sufficient to confidently identify variation unique to an individual in a single run. Other production-level systems, such as those from Pacific Biosystems (https://www.pacb.com/) and Oxford Nanopore (https://nanoporetech.com/) Technologies, produce significantly fewer reads per run than the Illumina systems. Generating long-read output, the reads they produce are generally 10 to 100 times the size, which increases their value for the assembly of complicated genomic or transcriptomic regions. However, long-read sequencing remains expensive for most purposes. At the time of writing, whole-genome sequencing at approximately 15 times coverage using short-read

technology can be generated at a core facility for approximately $500 per individual. Overall, these production systems increase sequencing capacity and improve accessibility for researchers by reducing sample costs so that NGS technology can be applied effectively to more research questions.

Benchtop systems represent the category of sequencers, which literally live in a laboratory on the benchtop. In general, they have moderate sequencing capacity (1.2–150 Gbp sequence data and 4–400 million reads per run) but represent an improvement in accessibility for investigators. These instruments serve smaller communities of researchers (or even a single laboratory) compared with the production systems. Therefore, benchtop systems often allow faster data generation because the researchers are not sharing the instrument with as many other users and do not have to wait as long to use the machine. Examples of these include the Illumina MiSeq, iSeq 100, the ThermoFisher Ion Gene Studio S5, and the Oxford Nanopore GridION. Benchtop systems are optimized for investigators who have smaller sample sizes (e.g., preliminary studies), who wish to perform transcriptome analyses, which generally require lesser sequence output, or who have targeted sequencing objectives. These systems can also be used by those who want to use sequencing in clinical medicine, although the applications for such sequencing are still developing.

Like much of technology, sequencers are becoming more efficient and are beginning to come in much smaller packages. Portable systems are designed to allow sequencing without the requirement of the support of a full laboratory. Examples of these systems are currently available from Oxford Nanopore and include the MinION and SmidgION (a small-capacity sequencer that can be operated with a smartphone). The amount of data generated is impressive but generally lower than either the benchtop or production systems. Both systems are supported by equally portable sample preparation and analysis tools. Possible applications of these mobile systems include stall-side diagnoses of an infectious pathogen or DNA verification of an individual's identity. Analyses can be conducted in a short amount of time to answer time-sensitive questions (e.g., what strain of a virus is present?).

Sequence Reads

Just as the number and type of available sequencers have proliferated, so too have the types of data produced. The primary distinction of sequence data is the length of reads generated by the sequencing instrument. There are 2 main categories for NGS data: short and long reads. Short sequence reads (short reads) are usually shorter than 500 bp in length, whereas long sequence reads (long reads) exceed 1000 bp.[20] Sanger sequencing reads are between these 2 classifications.

Short-read sequence is the most common type of NGS data reported in the literature. The main advantage of short-read sequencing is that a single instrument can produce large amounts of data with high-quality base calls in 1 run (see Table 1). This ability gives researchers/clinicians options for their sequencing experiments: they can generate high levels of coverage on a few individual samples to support identification of sequence variants in DNA and characterization of gene expression (Table 2), or they can pool samples to efficiently and cost-effectively generate data for large sample sets. Short reads can be classified as either single-end or paired-end. This designation refers to whether sequence was

generated from both or just 1 end of the captured DNA fragments. Single-end short reads are valuable for rapid and inexpensive characterization of DNA sequence or gene expression. However, their use is limited for the characterization of complex sequence regions such as sequence repeats or alternative splicing. This limitation results from ambiguity in aligning the single-end read back to a reference genome. The advantage of paired-end reads is that sequence from both ends of a DNA fragment of known length is generated. Then, information from both ends of the read can be used in parallel, which enhances the strategies used to address characterization of complex sequences. Paired-end reads, which align to the reference genome at a distance (between the reads) less than or greater to what was expected, can indicate the presence of a sequence variant such as an insertion or deletion, or, in the case of transcriptome data, can reveal patterns of alternative splicing. Figure 2 shows both single-end and paired-end short reads and the application of those reads to the characterization of DNA and RNA sequences.

Table 2. Summary of various next-generation sequencing application categories, including the type of variants it is possible to assay and how much sequence data are required

Target	Category	Application/Detection	Coverage/Sequence Required
DNA	Genome sequencing	SNPs, INDELs, CNVs, genotyping	10× to 60× coverage
	Exome sequencing		100× coverage
	Targeted sequencing	SNPs	15–100 million reads
	Methylation	ChIP, SNPs, chromosome conformation	15–30 million reads
	De novo assembly		140× coverage
		Bisulfite	
		SNPs, INDELs, CNVs, genotyping	
RNA	Transcriptome sequencing	Differential expression, small RNAs, alternative splicing	10–100 million reads
	Targeted Sequencing	CLIP, transcript panels, tag capture	5–40 million reads
	De novo assembly		> 100 million reads
		Differential expression, small RNAs, alternative splicing	

Abbreviations: ChIP, chromatin immunoprecipitation; CLIP, cross-linking immunoprecipitation; CNVs, copy number variations; INDELs, insertions/deletions; SNPs, single nucleotide polymorphisms

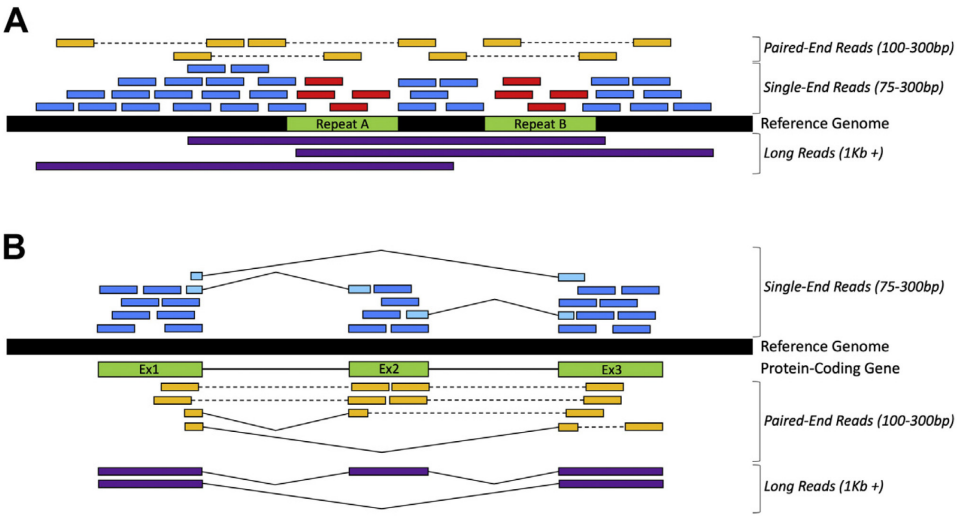

Figure 2. Various read types and read lengths and the application of those reads for DNA and RNA sequencing. In both cases, the reads are aligned to a reference genome (*black rectangle*) for analysis. (**A**) DNA sequencing: the region of the genome depicted contains 2 copies of a repeated motif (*green rectangles*). Single-end short reads are aligned across the genome at unique locations (*blue rectangles*) or multiple locations (*red rectangles*) if they originated from a repeat sequence. Paired-end short reads (*yellow rectangles joined by dashed lines*) can help to characterize the repeat regions because they align to the repeats and are anchored by alignment to unique sequences. Long reads (*large purple boxes*) align uniquely to the reference genome and can be used to characterize repeat sequence because they span the entire region. (**B**) RNA sequencing: the area of the genome depicted encodes a protein-coding gene (*green boxes connected by solid lines*). Single-end short reads map to sequence representing the exonic regions of the gene and can be mapped with a gapped alignment (*light blue rectangles joined by angled solid lines*) representing the union of 2 exons by splicing. Paired-end short reads also align to the exonic regions of the gene and can be used to define exon order in a transcript by linking multiple exons together. Long reads can help determine full-length transcripts and can be used to separate overlapping transcript structures.

Long-read sequence data are increasing in popularity for NGS experiments. The instruments that generate these data generally produce fewer sequence reads per run, but the reads they do provide are significantly longer than those from any short-read NGS platform (see Table 1). When first released, reads from these instruments averaged 1100 bp in length. Improvements in chemistry quickly increased the expected read length to 10,000 bp, with some reads spanning 60,000 bp. Genome assembly and the investigation of large-scale structural variation is aided by long-read sequencing because the long reads can sometimes span the length of repetitive regions of the genome, or moderately sized insertions/deletions. The read length achievable has led to the preferential use of this platform for genome assembly and scaffolding, as was the case in the newest assembly of the equine genome, EquCab3.[10] Long-read sequencing has also helped to resolve the structure and sequence of highly repetitive regions such as the equine major histocompatibility complex.[21]

Long sequence reads can also be used for annotation of alternative splicing in the transcriptome because the long reads can span entire transcripts. Figure 2 shows how long reads can be used in both DNA and RNA sequencing (RNA-seq) applications. A common strategy is to combine both short-read and long-read data in a single NGS experiment to exploit the advantages offered by each.

Applications

With production instruments now capable of producing terabytes of data each run, the sequencing of a horse's entire genome is now arguably the most common use of NGS technologies. This type of sequencing can allow the identification of inherited or de novo variation associated with disease. Whole-genome sequencing in the horse has enabled the discovery of variation that can be assayed to identify the risk of disease or for use in diagnosis. Some findings that resulted from the use of whole-genome sequence include missense variants causative of lavender foal syndrome,[22] immune-mediated myositis,[23] the identification of a locus associated with risk for squamous cell carcinoma in Haflingers,[24] a nonsense mutation associated with hydrocephalus in Friesians,[25] a splice-site mutation in Friesian horses with dwarfism,[26] and a large deletion associated with occipitoatlantoaxial malformation.[27] A practical alternative to whole-genome sequencing can be sequencing of only the exome, the regions of the genome that code for the exons of protein-coding genes. This approach requires that the exonic sequence is captured (either in solution or on an array) to prepare the DNA for sequencing. Because the region to be sequenced is reduced relative to the whole genome, this approach can enable a researcher to generate sequence from a larger number of individuals. The primary limitation in horses is the availability of capture technology. Exome sequencing has been used in the horse to identify variants relative to racing performance in quarter horses.[28] In each of these examples, the sequence generated was aligned to the reference genome and variants differing between affected horses and the reference sequence, or compared with healthy controls, were identified. As part of this process, after variants that either fit the hypothesized mode of inheritance or are found in candidate genes are identified, the possible function of each can be predicted using the genome annotation. In cases where genes may not be annotated, the region can be aligned to orthologous loci of other species. The impact of genomic variants on gene expression can also be assayed through RNA sequencing of the appropriate tissues.

The process of sequencing of the transcriptome (any portion of the DNA actively being transcribed into RNA at the time the tissue is sampled) is similar to that of sequencing DNA. The exception is an initial reverse transcription step, through which the isolated RNA is converted to double-stranded, copy DNA. The library preparation method used for RNA-seq depends on the question at hand. Poly-A^+ selected libraries capture most messenger RNA and some long noncoding RNA, as long as a poly-A tail is present on the transcript. Poly-A^+ library preparation and paired-end sequencing are the most common means to assess the expression of protein-coding loci. Differential expression can also be assayed using 3′ tag-seq (Quantseq; Lexogen, Greenland, NH), a method in which libraries are created for only the 3′ end of each RNA molecule present in the sample. Tag-seq does

not allow the identification of gene isoforms, but, by focusing sequencing efforts on only the terminal end of each transcript, differential expression analyses require significantly lower sequencing depth (~6 million reads per sample)[29,30] using single-end reads, therefore reducing overall cost. As in whole-genome sequencing, reads from RNA-seq are mapped to the reference genome or, in some cases, the transcriptome. In the horse, RNAseq data have been used to develop and improve gene annotation.[31–34] The relative abundance of each transcript can then be quantified using the available gene annotation and compared between treatments or disease states.[35–39] The sequencing of mRNA through poly-A$^+$ selection not only allows the quantification of each transcript but can provide insight into splice-site variation. Further, RNA libraries are often stranded, meaning the sequence generated distinguishes the strand of DNA from which the transcript was derived. This technique is a powerful method to identify and distinguish antisense transcripts. The advent of long-read technology can also be applied to studies of the transcriptome. Iso-seq is the use of PacBio sequencing, enabling the profiling of full-length RNA transcripts.[40] This methodology reduces 3′ sequencing bias, which is common in poly-A$^+$ library preparation, and is a powerful means to annotate genomes and identify variation in codon usage. However, as a long-read technology, Iso-seq is thus far too expensive for most clinical investigations. In contrast, Poly-A$^+$ library preparation neglects sequencing of small RNAs such as microRNAs, which can be assayed with a special, small RNA library preparation method. MicroRNAs are small (21–25 nucleotides) RNA fragments encoded by the animal's genome. Although they do not function to create proteins, they can bind to and silence the expression of protein-coding genes; therefore, their activity in posttranscriptional modification can significantly affect genome function.[41] In horses, microRNA profiles have been proposed as useful biomarkers for infection[42] or other disorders.[43,44]

In addition to the identification of genomic variants and the transcript expression, NGS can be used to understand chemical modifications to the DNA, such as methylation, or to identify regions of the genome interacting with protein. DNA methylation, a chemical modification of cytosine to 5-methylcytosine, is a common epigenetic mechanism involved in silencing gene expression.[45] Although the inheritance of some epigenetic modifications, such as DNA methylation, is not completely understood, like RNA-seq, examining methylation patterns can help to understand differences in gene regulation and expression between diseased and healthy individuals. Similar to chromatin immunoprecipitation (ChIP; discussed later), genome-wide methylation can be assayed by using antibodies to precipitate DNA having 5-methylcytosine modifications; that DNA is then sequenced on a next-generation platform (MeDIP-seq).[46] To the authors' knowledge, this method has not yet been applied in a case of equine disease research; however, this technique has been used to characterize changes in genomic methylation in equine skeletal muscle caused by exercise.[47,48]

ChIP is a method by which regions of the DNA involved in an interaction with protein are isolated.[49] Those regions of DNA can then be sequenced using standard next-generation methodology (ChIP-seq), and the resulting DNA fragments aligned to the reference genome to identify genomic regions involved in the interaction. Similarly, cross-linking immunoprecipitation (CLIP) is a method that enables the isolation of RNA transcripts specifically interacting with a protein.[50] The captured transcripts can be sequenced (CLIP-seq) to identify regulatory aspects of gene expression. Sequence variants in the regions of interaction

can alter binding efficiency and thus function. In addition, alteration in protein-DNA or protein-RNA interactions can uncover functional information regarding molecular mechanisms of disease. Therefore, these methods can be used to investigate both the functional significance of genomic variation as well as to identify alterations in genome activity and sequence composition associated with a treatment or disease. Of note, as is the case in transcriptome sequencing, these methods of capturing information on genome function only reveal information about the genome's activity within the tissue or cell population sampled at time of sampling.

Data Generation and Handling

For most any platform, the process of sequencing is similar. The general workflow of an NGS experiment is presented in Figure 3. The goals for interpretation and application of the data generated can result in alterations of this general approach. The sample necessary depends on the question at hand. For gene expression, RNA must be isolated from a tissue relevant to the phenotype of interest. Because RNA is relatively unstable, care has to be taken to either preserve the sample in an RNA-stabilization solution (e.g., RNAlater, Sigma-Aldrich, St Louis, Missouri; DNA/RNA Shield, Zymo Research, Irvine, California) or the tissue must be flash-frozen immediately after collection until processing. ChIP data can also be derived from a flash-frozen sample or from samples subjected to a cross-linking protocol, commonly performed with formaldehyde, at the time of collection. If the goal is to identify genomic variation, genomic DNA must first be isolated from a sample of the individual. Blood and tissue are commonly used samples for DNA isolation, although hair follicles can also produce adequate DNA for sequencing of target genes or the whole genome. The isolated DNA is hydrolyzed or sheared to create fragments of similar size and is processed for library preparation. Methods for library preparation are conceptually similar, although there are platformspecific processes to make the input nucleic acid ready for sequencing on a particular instrument. Barcode sequences can be included and allow individual samples to be pooled in a single sequencing run.

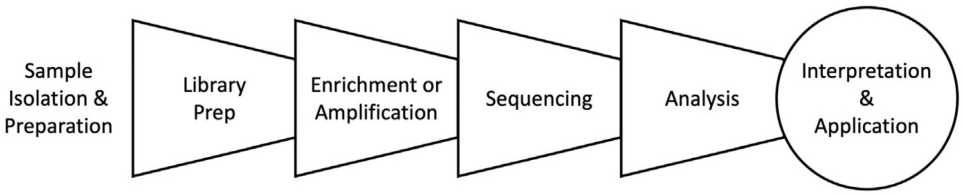

Figure 3. General workflow of an NGS experiment.

The general features of NGS data analysis are similar regardless of the platform used to generate the data (Fig. 4). The data received from most sequencing methods are in the form of fastq files. These files encode both the sequence identity of each read as well as an associated quality metric. Data processing then involves an initial step of quality control where any adapter sequences necessary for library preparation are removed, and the

sequence is also trimmed to eliminate base calls that do not meet a designated quality threshold. It is standard for the 3′ end of each read to be of lesser quality than the 5′ end, and thus much of the trimming occurs on this portion of the read. If a paired-end library is sequenced, it is important that the data from the 2 ends of each pair remain associated; if 1 read is completely removed because of poor quality, its paired read must also be removed from the dataset. Once the data are preprocessed for quality control, the reads are aligned with the equine reference genome, or possibly the transcriptome (in the case of RNA-seq efforts). This process is computationally expensive and, depending on available computing resources and amount of data being processed, it could take days to weeks. However, the aligned reads (in bam files) can be visualized in software such as the Integrative Genomics Viewer[51–53] or JBrowse.[54]

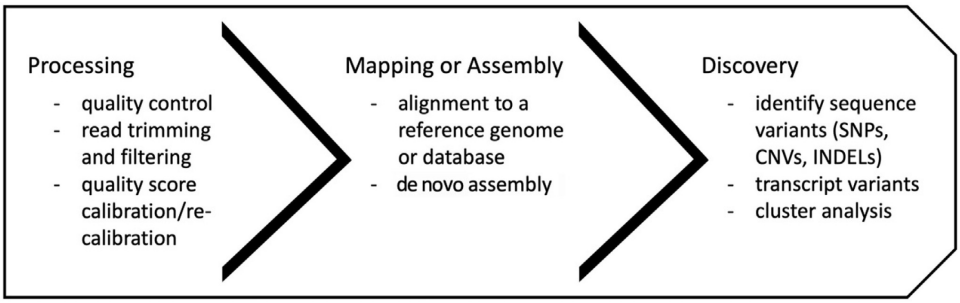

Figure 4. General features of NGS data analysis. CNV, copy number variation; INDEL, insertion/deletion; SNPs, single nucleotide polymorphisms.

Once the sequencing reads are aligned, variants within the newly sequenced individual and between it and the reference genome can be identified. Variant calling identifies single nucleotide polymorphisms (SNPs), insertions/deletions, or structural variation that differ between each individual and the reference genome, within an individual (i.e., heterozygous sites), or between study individuals. A variety of variant calling software is available and has been reviewed elsewhere.[55,56] The choice of a variant caller depends to some extent on the question asked (e.g., rare variant identification vs. population frequency). In addition to the selection of variant calling software, the quality of the output also depends on the depth of sequence coverage, sequence quality, and ability to filter false-positive signals. Once variants are identified, the genome annotation provides a means to predict the functional impact (e.g., nonsynonymous mutation or splice-site variant) of each. For cases with apparent simple inheritance, several databases exist to help identify candidate genes, or genes previously associated with similar phenotypes. Online Mendelian Inheritance in Man (OMIM; https://www.omim.org/), and Online Mendelian Inheritance in Animals (OMIA; https://omia.org/home/) provide information on thousands of known mendelian traits. Previously annotated variants and quantitative trait loci (loci associated with complex disease) are also often cataloged and can serve as valuable resources when investigating putative functional variation; these are available in databases such as the European Variation Archive (https://www.ebi.ac.uk/eva/). However, not all variants or genes with an essential

physiologic function are annotated as such. In contrast, not all loci computationally predicted or modeled to affect gene or protein function necessarily do so. Validation of function of a variant requires significant subsequent work beyond their discovery.

For transcriptomic data, variants can be called in a manner similar to that used for whole-genome sequencing. However, the purpose of RNA-seq is often not to identify variation but altered expression of gene expression between affected versus unaffected tissues. Differential expression analyses are conducted based on the quantification of reads observed per transcript. Data must first be normalized to account for differences in sequencing depth, and, depending on the method used, analyses may also consider transcript length. Reviews of methods for quantification and differential expression analyses of RNA-seq data outline the statistical models used and assumptions underlying each approach.[57,58] Transcriptomic data are often used to investigate the function of putative causative variants or to identify gene pathways associated with disease, such as in the case of stationary night blindness of Appaloosas[35] and Arabian cerebellar abiotrophy.[59]

Validation of Results

Even though NGS is becoming common, there are currently no standards set forth by the veterinary industry on the interpretation or use of DNA sequencing or RNA-seq data. Therefore, in the use of genetic or genomic information for equine management the onus is on researchers, clinicians, owners, or other end users to evaluate the process by which the data were discovered and validated. In human medicine, various interest groups, such as the Next-Generation Sequencing: Standardization of Clinical testing II informatics workgroup[60] and the American College of Medical Genetics and Genomics,[61] have worked to address the means to ensure that rigorous standards of variant discovery and validation are met. Some of the principles put in place by these entities include outlining a vocabulary useful to classify variant function and assist clinicians in using information regarding genetic tests in practice.[60,61] Another idea shared by both groups emphasizes that variant function and predictive ability need to be validated in individuals unique to those used in the discovery process and the variant frequency within the population (e.g., breed in the case of horses) should be described. Toward a similar goal of standardizing how genomics research is implemented, validated, and applied, the international equine genomic research community recently put forth a "Consensus Statement on the Translation and Application of Genomics in the Equine Industry" (Havemeyer Principles 2019: https://horsegenomeworkshop.com/values).[62] In this statement, the researchers acknowledge that genomics and discovery using NGS holds significant promise to improve equine well-being. However, with the complexity of disease and of genome function, the community agreed the most significant benefit of genomics to the horse lies in discovery that encompasses several key elements. These elements include ensuring that genomic research is reproducible and peer reviewed, ethical, and performed and communicated with transparency. As the use of NGS increases, these guidelines will need to become more clearly defined because, although the potential for genomics to improve equine health and well-being is undeniable, its successful application also depends on the rigor of the research behind the discoveries.

Future Directions

The degree of advancement of genomic tools for researchers and clinicians in the past 10 years has been tremendous. The accelerated rate of discovery is likely to continue, and, with decreasing costs of NGS, the use of this method in the diagnosis, prevention, and management of disease is likely to become common practice. As researchers build a better understanding of how genetic variation alters an individual's ability to respond to treatment or optimize performance, the idea of personalized or precision management for horses is far reaching. In addition, a greater understanding of genomic relationships among individuals, as well as how genomic variation contributes to complex phenotypes such as disease, lends itself to use in genomic selection, or the incorporation of genomic information with phenotype data to predict an animal's breeding value for a trait or traits of interest. The improved understanding of genome function and disease susceptibility supported by the application of NGS can lead to better horse health and welfare.

Disclosure – The authors have nothing to disclose.

References

1. Wade CM, Giulotto E, Sigurdsson S, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 2009;326:865–7.
2. McCue ME, Bannasch DL, Petersen JL, et al. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. PLoS Genet 2012;8:e1002451.
3. Schaefer RJ, Schubert M, Bailey E, et al. Developing a 670k genotyping array to tag w2M SNPs across 24 horse breeds. BMC Genomics 2017;18:565.
4. Rudolph JA, Spier SJ, Byrns G, et al. Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding. Nat Genet 1992;2:144–7.
5. Santschi EM, Purdy AK, Valberg SJ, et al. Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses. Mamm Genome 1998;9:306–9.
6. Chowdhary BP. Equine genomics. College Station (TX): Wiley-Blackwell; 2013.
7. Ghosh S, Qu Z, Das PJ, et al. Copy number variation in the horse genome. PLoS Genet 2014; 10:e1004712.
8. Glaser KE, Sun Q, Wells MT, et al. Development of a novel equine whole transcript oligonucleotide GeneChip microarray and its use in gene expression profiling of normal articular-epiphyseal cartilage. Equine Vet J 2009;41:663–70.
9. Mienaltowski MJ, Huang L, Stromberg AJ, et al. Differential gene expression associated with postnatal equine articular cartilage maturation. BMC Musculoskelet Disord 2008;9:149.
10. Kalbfleisch TS, Rice ES, DePriest MS Jr, et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol 2018;1:197.
11. Putnam NH, O'Connell BL, Stites JC, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 2016;26:342–50.

12. Belton JM, McCord RP, Gibcus JH, et al. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 2012;58:268–76.
13. Burns EN, Bordbari MH, Mienaltowski MJ, et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet 2018;49:564–70.
14. Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. Nature 2017;550:345–53.
15. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc 2017;12: 213–8.
16. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463–7.
17. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26: 1135–45.
18. Metzker ML. Sequencing technologies the next generation. Nat Rev Genet 2010;11:31–46.
19. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51.
20. Junemann S, Sedlazeck FJ, Prior K, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol 2013;31:294–6.
21. Viluma A, Mikko S, Hahn D, et al. Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology. Sci Rep 2017;7:45518.
22. Brooks SA, Gabreski N, Miller D, et al. Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for Lavender Foal Syn-drome. PLoS Genet 2010; 6:e1000909.
23. Finno CJ, Gianino G, Perumbakkam S, et al. A missense mutation in MYH1 is associated with susceptibility to immune-mediated myositis in Quarter Horses. Skelet Muscle 2018;8:7.
24. Bellone RR, Liu J, Petersen JL, et al. A missense mutation in damage-specific DNA binding protein 2 is a genetic risk factor for limbal squamous cell carcinoma in horses. Int J Cancer 2017; 141:342–53.
25. Ducro BJ, Schurink A, Bastiaansen JW, et al. A nonsense mutation in B3GALNT2 is concordant with hydrocephalus in Friesian horses. BMC Genomics 2015; 16:761.
26. Leegwater PA, Vos-Loohuis M, Ducro BJ, et al. Dwarfism with joint laxity in Friesian horses is associated with a splice site mutation in B4GALT7. BMC Genomics 2016;17:839.
27. Bordbari MH, Penedo MCT, Aleman M, et al. Deletion of 2.7 kb near HOXD3 in an Arabian horse with occipitoatlantoaxial malformation. Anim Genet 2017;48: 287–94.
28. Pereira GL, Malheiros JM, Ospina AMT, et al. Exome sequencing in genomic regions related to racing performance of Quarter Horses. J Appl Genet 2019;60: 79–86.
29. Meyer E, Aglyamova GV, Matz MV. Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Mol Ecol 2011;20:3599–616.
30. Lohman BK, Weber JN, Bolnick DI. Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. Mol Ecol Resour 2016;16:1315–21.
31. Coleman SJ, Zeng Z, Wang K, et al. Structural annotation of equine protein-coding genes determined by mRNA sequencing. Anim Genet 2010;41(Suppl 2):121–30.
32. Coleman SJ, Zeng Z, Hestand MS, et al. Analysis of unannotated equine transcripts identified by mRNA sequencing. PLoS One 2013;8:e70125.

33. Hestand MS, Kalbfleisch TS, Coleman SJ, et al. Annotation of the Protein Coding Regions of the Equine Genome. PLoS One 2015;10:e0124375.
34. Mansour TA, Scott EY, Finno CJ, et al. Tissue resolved, gene structure refined equine transcriptome. BMC Genomics 2017;18:103.
35. Bellone RR, Holl H, Setaluri V, et al. Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. PLoS One 2013;8:e78280.
36. Tessier L, Cote O, Clark ME, et al. Impaired response of the bronchial epithelium to inflammation characterizes severe equine asthma. BMC Genomics 2017; 18:708.
37. Pacholewska A, Jagannathan V, Drogemuller M, et al. Impaired Cell Cycle Regulation in a Natural Equine Model of Asthma. PLoS One 2015;10:e0136103.
38. Valberg SJ, Perumbakkam S, McKenzie EC, et al. Proteome and transcriptome profiling of equine myofibrillar myopathy identifies diminished peroxiredoxin 6 and altered cysteine metabolic pathways. Physiol Genomics 2018;50:1036–50.
39. Finno CJ, Bordbari MH, Valberg SJ, et al. Transcriptome profiling of equine vitamin E deficient neuroaxonal dystrophy identifies upregulation of liver X receptor target genes. Free Radic Biol Med 2016;101:261–71.
40. Wang B, Tseng E, Regulski M, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 2016;7: 11708.
41. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. NatRev Genet 2004;5:522–31.
42. Cowled C, Foo CH, Deffrasnes C, et al. Circulating microRNA profiles of Hendravirus infection in horses. Sci Rep 2017;7:7431.
43. Barrey E, Bonnamy B, Barrey EJ, et al. Muscular microRNA expressions in healthy and myopathic horses suffering from polysaccharide storage myopathy or recurrent exertional rhabdomyolysis. Equine Vet J Suppl 2010;(38):303–10.
44. Desjardin C, Vaiman A, Mata X, et al. Next-generation sequencing identifies equine cartilage and subchondral bone miRNAs and suggests their involvement in osteochondrosis physiopathology. BMC Genomics 2014;15:798.
45. Razin A, Riggs AD. DNA methylation and gene function. Science 1980;210: 604–10.
46. Weber M, Davies JJ, Wittig D, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 2005; 37:853–62.
47. Gim JA, Hong CP, Kim DS, et al. Genome-wide analysis of DNA methylation before and after exercise in the thoroughbred horse with MeDIP-Seq. Mol Cell 2015;38:210–20.
48. Lee JR, Hong CP, Moon JW, et al. Genome-wide analysis of DNA methylation patterns in horse. BMC Genomics 2014;15:598.
49. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 2012;13:840–52.
50. Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 2008;456:464–9.
51. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6.

52. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
53. Robinson JT, Thorvaldsdottir H, Wenger AM, et al. Variant Review with the Integrative Genomics Viewer. Cancer Res 2017;77:e31–4.
54. Buels R, Yao E, Diesh CM, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016;17:66.
55. Bohannan ZS, Mitrofanova A. Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables. Comput Struct Biotechnol J 2019;17:561–9.
56. Li Z, Wang Y, Wang F. A study on fast calling variants from next-generation sequencing data using decision tree. BMC Bioinformatics 2018;19:145.
57. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One 2017;12:e0190152.
58. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics 2015;14: 130–42.
59. Scott EY, Woolard KD, Finno CJ, et al. Variation in MUTYH expression in Arabian horses with Cerebellar Abiotrophy. Brain Res 2018;1678:330–6.
60. Gargis AS, Kalman L, Bick DP, et al. Good laboratory practice for clinical nextgeneration sequencing informatics pipelines. Nat Biotechnol 2015;33:689–93.
61. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17:405–24.
62. Bailey E, Finno C. Translation and application of equine genomics: The Have-meyer principles. Equine Vet J 2019;51:273.