ORIGINAL PAPER   OPEN ACCESS

Prospects of Whole Genome Sequencing in Animal Breeding

[1]Aakanksha Rathore,[2] Ramesh Kumar Singh and [3] Rajesh Kumar

[1]Veterinary Assistant Surgeon, Goat Breeding Farm, Pakariya, Gaurela-Pendra-Marwahi, C.G. (India)
[2] Assistant Professor, Department of Animal Genetics and Breeding, BVC, Patna
[3] Assistant Professor (A.H.), Department of Agronomy, BAC, Sabour

*Corresponding author: Ramesh.kumarvet@gmail.com

T he development of high-throughput sequencing technologies has revolutionized animal genetics and genomics. The widespread use of whole-genome sequencing (WGS) allows detection of a full range of common and rare genetic variants of different types across almost the entire genome which facilitates rare disease research and clinical applications. Further, these may improve common disease discovery and annotation of the causal variants. WGS may be a predominant technology for genetic analysis. This is a fundamental change compared to previous decades of animal and human genetic studies that have relied on genetic markers that are indirect proxies of other genetic variants in the surrounding region or sequencing data only from the exonic regions of the genome. Functional interpretation of variants discovered by WGS is an important component of animal genetics studies and is essential for revealing the effects of variants on traits. Genome-wide functional genomics assays now allow for increasingly accurate detection, characterization, and prediction of the molecular effects of variants. However, these effects reflect the full complexity of genome function, our understanding of which is incomplete, much remains to be discovered regarding variant molecular effects and their potential for impacting higher-level organismal phenotypes. The variants are discovered by WGS and genome-wide functional genomic approaches used for analysis of functional effects of these variants. These are foundational building blocks for the discovery and interpretation of genetic effects on animal production traits

WGS TECHNOLOGIES

The first aim of a typical WGS study is to create a high-quality map of genome variation for the samples of interest. This crucial step lays the foundation for all downstream analyses aimed at genome interpretation and genetic discovery.

The methods used to map genome variation depend heavily on the sequencing technology and depth of coverage obtained. There are currently three general WGS strategies.

(1)   short-read WGS using the Illumina technology, which currently yields paired-end 150 bp reads with low error rates in the range of 0.1%–0.5%;

(2)   long-read WGS using single-molecule technologies from Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT), which yield 10–100 kb reads—and occasionally much longer—with high error rates in the range of 10%–15%;

(3)   linked-read WGS us-ing the technology from 10X Genomics, which generates barcoded Illumina short-reads from longer molecules (e.g.,50 kb).

To distinguish variants from errors, each base in the genome must be sequenced multiple times from randomly sampled DNA molecules. Deeper coverage improves variant detection sensitivity and also improves accuracy by allowing for more sophisticated filtering schemes. The most current WGS studies are employing deep WGS (>20x) in practice. Since high-quality *de novo* assembly is not possible from short reads, standard WGS analysis pipelines align reads to the reference genome and map variants relative to the reference. Most modern pipelines use BWA-MEM for alignment and a combination of tools for subsequent processing. This issue is especially trouble-some for large-scale trait association studies where subtle genotyping biases can grossly inflate false positives and where reprocessing of large datasets to achieve harmonization would require much time and expense. Data compatibility is also extremely important for small-scale studies that aim to accurately compare variant calls with public data-bases such as gnomAD. A recent multicenter effort established a model for implementing functionally equivalent pipelines alleviate batch effects and enable data sharing. The followings are the classes of variants detected after WGS.

Single-Nucleotide Variants and Small Insertion/Deletion Variants

Single-nucleotide variants (SNVs) and small insertion/deletion variants (indels) (<50 bp) comprise the vast majority of variants. There are 3–4 million SNVs and 0.4–0.5 million indels apparent in a typical comparison of one genome sequence versus the reference. The vast majority of this huge number of variants have no functional impact at the molecular or phenotypic level, every genome has >100 protein truncating variants (PTVs) that introduce a premature stop codon. Nonsynonymous or missense SNVs or in-frame indels lead to amino acid changes, which can be entirely benign or cause a severe disease. Finally, these variants can affect gene regulation by affecting transcriptional and posttranscriptional regulatory elements. Fundamentally, for an SNP or small indel to have an effect on gene regulation, a sequence-specific regulator whose

activity is differentially affected by the two alleles is needed—at least at some point during development. These include transcription and splicing factors that bind to specific DNA motifs, as well as noncoding regulatory RNAs such as miRNAs. These small variants are the easiest class of variants to detect from short-read data. In general, SNV/indel detection algorithms scan the reference genome in search of collections of aligned reads that exhibit mismatches, insertions, or deletions in a manner that suggests germline variation rather than sequencing or alignment error. Existing widely used tools are highly effective in the 72% of the genome that is unique and allows for accurate read alignment, with levels of sensitivity and specificity that exceed 99.5% for SNVs and 95% for indels.

STRUCTURAL VARIATION

Structural variation (SV) is a diverse form of genome variation R50 bp in size that includes copy number variants (CNVs), rear-rangements, and mobile element insertions (MEIs). SVs are few in number compared to SNVs and indels but have more severe consequences on average due to their size. SVs can exert functional effects by changing gene dosage, disrupting gene function (similar to PTVs), or rearranging regulatory elements and/or genes to alter genomic context. Unsurprisingly, extremely large variants that delete or duplicate many genes or even entire chromosomes typically have drastic phenotypic effects and are not observed in most individuals. Smaller and more prevalent forms of SV typically affect only one or a few genes or lie within noncoding regions. Although SVs account for merely 0.2% of total variants, recent WGS-based studies have estimated that they account for 3%–7% of common variants with *cis*-acting effects on gene expression, a much larger fraction of rare expression-altering variants, and 4%–12% of high-impact coding alleles. SV is recognized to be the most difficult form of variation to detect reliably from short-read data.