REVIEW ARTICLE DOI: 10.sjas.202510060006

Current Trends, Challenges, and Optimization Strategies in Bioinformatic Pipelines for Whole Genome Sequencing of Non-Model Species

Javeria Ayub [1], Sara Bibi [2], Farwa Jabbir [3], Maaz Ullah [4], Hafiz Ishfaq Ahmad [5]*

[1] Department of Zoology, The Islamia University of Bahawalpur, Pakistan.
[2] Department of Animal Science, University of Sargodha, Pakistan.
[3] Department of Biotechnology, University of Sargodha, Pakistan.
[4] College of Animal Science and Technology, Northwest A&F University, Yangling, Xianyang, Shaanxi, China.
[5] Department of Animal Breeding and Genetics, Faculty of Veterinary and Animal Sciences, The Islamia University of Bahawalpur, Pakistan.

*Corresponding author: Hafiz Ishfaq Ahmad
Email: ishfaq.ahmad@iub.edu.pk

Cite this Article: Ayub J, Bibi S, Jabbir F, Ullah M, Ahmad HI (2026). Current trends, challenges, and optimization strategies in bioinformatic pipelines for whole genome sequencing of non-model species. SciNex Journal of Advanced Sciences. 01 (01): 202510060006

ABSTRACT: Whole genome sequencing (WGS) has become a central tool in evolutionary biology, conservation genetics, and agricultural genomics, enabling high-resolution analyses of genetic variation across diverse taxa. However, the application of WGS to non-model species presents substantial bioinformatic challenges, including incomplete or biased reference genomes, high levels of genetic diversity, variable sequencing depth, and limited computational resources. These constraints complicate pipeline design, variant discovery, and biological interpretation, particularly in agriculturally relevant systems where genomic outputs must be translated into practical outcomes.

This review critically examines current bioinformatic pipelines used for whole genome sequencing analyses in non-model species, with a focus on methodological trade-offs, sources of bias, and context-dependent optimization strategies. We synthesize recent advances in read processing, alignment and assembly approaches, variant calling frameworks, and functional annotation tools, and compare commonly used pipelines with respect to their suitability for non-model and agricultural applications. In addition, we highlight persistent limitations in benchmarking, reproducibility, and data integration, and discuss emerging trends such as long-read sequencing, pangenome frameworks, and machine learning-assisted pipeline optimization.

By integrating conceptual frameworks, comparative evaluations, and applied examples from crop, livestock, and pathogen genomics, this review provides practical guidance for designing robust and reproducible WGS bioinformatic workflows. The insights presented here aim to support informed methodological decision-making and to facilitate the effective translation of genomic data into agricultural improvement, conservation management, and biological discovery in non-model systems.

KEYWORDS: Whole genome sequencing; Non-model species; Pipeline optimization; Variant calling tools; Reference genome bias; Population genomics.

INTRODUCTION

The rapid evolution of high-throughput sequencing (HTS) technologies has democratized access to whole genome sequencing, moving it from a monumental undertaking for a few model organisms to a feasible tool for exploring the genetic fabric of virtually any species on Earth. This shift has been particularly transformative for the study of non-model species,organisms that lack the extensive genomic resources, such as high-quality reference genomes, detailed annotations, and established laboratory protocols, that are available for models like humans, mice, or fruit flies. Whole genome resequencing (WGR) of populations of non-model species offers unprecedented power to address fundamental questions in ecology, evolution, and conservation biology, from identifying genetic adaptations to climate change and deciphering demographic history to assessing the genomic basis of inbreeding and disease susceptibility in endangered populations (Hohenlohe et al., 2021; Leroy et al., 2021). However, the immense potential of WGR in non-model species is coupled with significant bioinformatic challenges. The analytical journey from raw sequencing reads to biological insight is fraught with complexities that are often exacerbated in non-model systems. Unlike work in model organisms, where standardized, validated pipelines exist, researchers working on non-model species must navigate a

labyrinth of decisions and potential pitfalls without a universal roadmap. The primary hurdle is the frequent absence of a high-quality, chromosome-level reference genome for the species of interest. While reference-free genome assembly is increasingly common, many studies rely on a draft genome or a reference from a closely related species. This can introduce substantial biases during read mapping, including increased rates of mismapping, lower mapping efficiency, and reference allele bias, where alleles present in the reference genome are artificially favored during variant calling (Günther & Nettelblad, 2019). These issues can systematically skew downstream population genetic analyses, such as estimates of nucleotide diversity, population structure, and selection signatures. The core of WGR analysis is the bioinformatic pipeline,a multi-step computational workflow that transforms raw sequencing data into a set of high-confidence genetic variants (e.g., single nucleotide polymorphisms - SNPs, indels). A standard pipeline typically involves quality control and adapter trimming, read alignment to a reference genome, processing of alignment files (e.g., duplicate marking, base quality recalibration), variant calling, and stringent variant filtering. Each step requires careful consideration of the software, parameters, and thresholds used, decisions that are highly sensitive to the specificities of the non-model study system (Shafer et al., 2017). For instance, the choice of aligner (e.g., BWA-MEM, Bowtie2) and its parameters can greatly affect mapping outcomes, particularly when dealing with divergent genomes. Similarly, variant callers like GATK's HaplotypeCaller or Samtools' mpileup, while powerful, are primarily optimized for human data and may require extensive parameter tuning and validation for non-model applications to balance the trade-off between sensitivity (finding all true variants) and specificity (avoiding false positives) (Poplin et al., 2018). Perhaps the most critical and often subjective stage is variant filtering. This process aims to remove spurious variant calls arising from sequencing errors, mapping artifacts, or misalignment. Researchers must filter based on a combination of quality metrics, such as read depth (DP), genotype quality (GQ), mapping quality (MQ), and strand bias. However, establishing appropriate thresholds is non-trivial. Overly stringent filtering may discard true, often rare or novel, variants, while overly lenient filtering will inundate the dataset with false positives that can invalidate subsequent analyses. In non-model species, the lack of known "true" variant sets or high-quality validation data makes it exceptionally difficult to benchmark filtering strategies and assess the true error rate of the final dataset (O'Neill et al., 2022). This has led to a concerning lack of standardization across studies, hindering reproducibility and meta-analyses. Beyond these core steps, the analysis of WGR data in non-model species often ventures into more complex territory. Many research questions require the identification of structural variants (SVs),larger genomic alterations like inversions, duplications, and translocations,which are even more challenging to accurately detect than SNPs, especially without a high-quality reference. Furthermore, the field is increasingly moving towards a population genomics approach, leveraging the full spectrum of genetic variation to infer demography, detect selection, and understand adaptive processes. This requires sophisticated methods for estimating allele frequency spectra, identifying runs of homozygosity (ROH) to measure inbreeding, and performing genome-wide association studies (GWAS) or genome-scans for selection (e.g., Fst, XP-CLR). Each of these analyses has its own assumptions and sensitivities to data quality and missingness, further compounding the analytical challenge. In applied and agricultural contexts, whole genome sequencing of non-model species plays an important role in crop improvement, livestock breeding, pathogen surveillance, and the conservation of agro-biodiversity. Many agriculturally important species lack high-quality reference genomes, making carefully optimized bioinformatic pipelines essential for translating genomic data into breeding, management, and disease-resistance strategies. This review article aims to provide a comprehensive overview and critical evaluation of bioinformatic pipelines for whole genome resequencing in non-model species. We will deconstruct the standard workflow, from raw data to variant calls, highlighting the key challenges and decision points at each stage. We will discuss best practices for quality control, read mapping, variant calling, and filtering in the context of divergent genomes and imperfect references. Furthermore, we will explore advanced applications, including structural variant detection and population genomic inference, and address the crucial issues of reproducibility and data management. By synthesizing insights from previous studies and emerging methodologies, this review seeks to serve as a practical guide for researchers navigating the complex yet powerful landscape of whole genome resequencing in the non-model world.

CHALLENGES IN WHOLE GENOME SEQUENCING FOR NON-MODEL SPECIES

To fully appreciate the complexities of bioinformatic pipeline design, it is first necessary to understand the fundamental challenges posed by non-model genomic systems. The application of whole genome sequencing (WGS) to non-model species has unlocked transformative potential in fields like evolutionary biology, conservation genetics, and ecology. However, the path to generating robust and biologically meaningful genomic insights is fraught with significant challenges that distinguish it from work in well-established model organisms. These challenges do not represent isolated technical issues but interact synergistically, often amplifying biases if pipeline design

and parameter selection are not carefully aligned with the biological characteristics of the study system. For example, in livestock genomics and crop wild relatives, reference bias can obscure adaptive loci associated with disease resistance or environmental tolerance, directly impacting breeding and selection decisions. These hurdles, which arise from the inherent biological characteristics of many non-model species and the resource-limited contexts in which they are often studied, can introduce substantial biases and errors if not carefully managed. This review synthesizes the primary challenges, focusing on the lack of reference genomes, high genetic diversity, issues of data quality and quantity, and daunting computational limitations. An overview of a typical whole genome sequencing bioinformatic workflow for non-model species is shown in Figure 1.

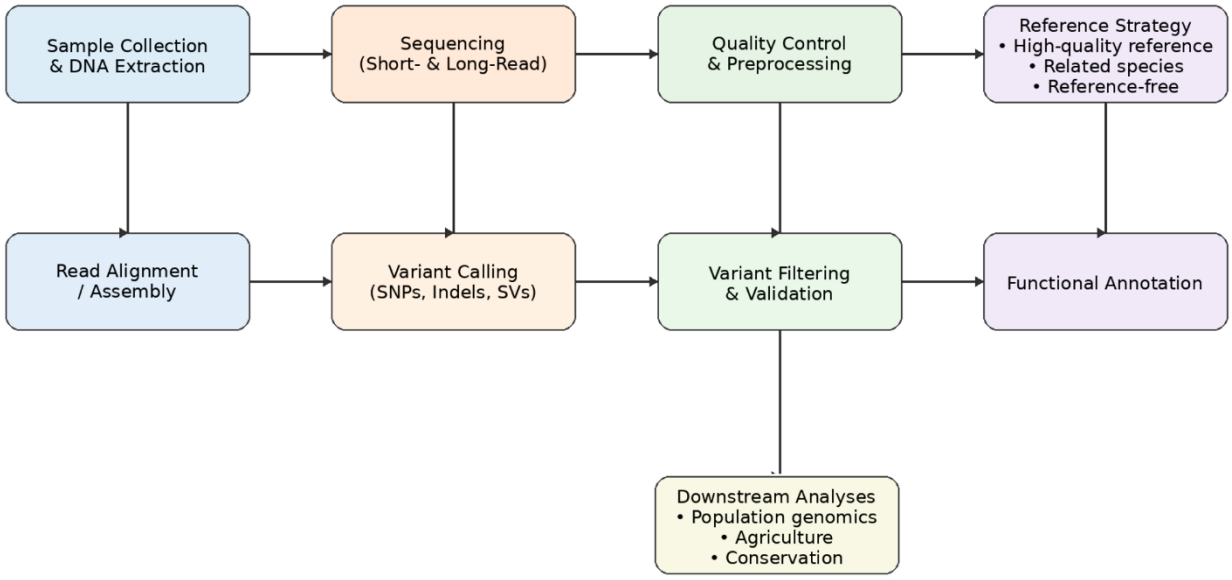

Figure 1: Conceptual workflow of whole genome sequencing bioinformatic pipelines for non-model species.

Lack of Reference Genomes

The cornerstone of most WGS analyses is a high-quality, chromosome-level reference genome. For non-model species, such a resource is frequently absent, presenting a fundamental and cascading set of problems. Many studies resort to reference-free genome assembly, a process that is itself highly challenging. High heterozygosity, repetitive elements, and polyploidy,common features in non-model species,can fragment assemblies and generate chimeric scaffolds. While long-read sequencing technologies (e.g., PacBio, Oxford Nanopore) have dramatically improved contiguity, achieving chromosome-level resolution often requires additional costly techniques like Hi-C or optical mapping, which are not always feasible (Hotaling et al., 2021). Consequently, many non-model species research relies on fragmented, incomplete draft genomes. These partial references create mapping biases during resequencing, as reads from unrepresented or misassembled regions may be discarded or mismapped, leading to a systematic loss of genetic variation and skewing downstream analyses (Günther & Nettelblad, 2019). A common alternative is the use of a reference genome from a closely related species. While practical, this approach introduces the problem of reference bias. During read alignment, sequences that diverge from the reference are less likely to map correctly or at all. This results in an artificial inflation of genetic similarity to the reference species and a severe under-calling of variants, particularly in the most divergent,and often most biologically interesting,genomic regions (Shafer et al., 2017). This bias directly impacts population genetic statistics, leading to underestimates of nucleotide diversity and distorted inferences of demographic history and selection, as the analysis becomes inherently biased towards conserved genomic areas.

High Genetic Diversity

Many non-model species, particularly those that are outcrossing or have large population sizes, exhibit levels of genetic diversity that far exceed those of classic model organisms. While a source of valuable information, this high diversity complicates bioinformatic procedures. Elevated heterozygosity poses a major challenge for reference-free assembly, often resulting in highly fragmented assemblies as haplotypes are assembled separately rather than merged into a single consensus sequence. For resequencing studies, high heterozygosity

increases the complexity of variant calling, as algorithms must distinguish true heterozygous sites from a background of sequencing errors and paralogous alignments. Furthermore, structural variations (SVs),including inversions, duplications, and large insertions/deletions,are abundant and poorly characterized in non-model species. Standard short-read aligners struggle to map reads accurately across breakpoints of SVs, leading to false positive variant calls and the misinterpretation of hemizygous regions as homozygous deletions. The detection of SVs requires specialized tools and often long-read sequencing data, adding another layer of analytical complexity and cost (Mahmoud et al., 2019). This high level of diversity and structural variation makes the alignment process less efficient and variant calling less accurate, ultimately obscuring the true patterns of genomic variation.

Data Quality and Quantity

Financial constraints often force researchers working on non-model species to make pragmatic decisions that impact data quality. Low-coverage sequencing (e.g., <10x coverage) is a common strategy to maximize the number of individuals sequenced within a budget. However, low coverage drastically increases genotype uncertainty. Distinguishing true heterozygous sites from sequencing errors becomes statistically challenging, often necessitating sophisticated genotype likelihood methods rather than direct genotype calling (Lou et al., 2021). While these methods enable valuable analyses like population structure inference, they are less powerful for detecting rare variants or performing association mapping, where accurate individual-level genotypes are critical. Moreover, WGS data is not free from technical artifacts. Sequencing errors are inherent to all platforms, and their profile differs between technologies (e.g., homopolymer errors in Nanopore, GC bias in Illumina). In non-model species, the absence of known true variants makes it difficult to calibrate base quality scores and empirically determine optimal filtering thresholds. PCR duplicates, caused during library preparation, can inflate coverage uniformity and must be identified and removed, but this process can mistakenly flag reads from paralogous regions as duplicates in the absence of a perfect reference, leading to the loss of genuine data. Managing these biases and errors without standardized resources requires careful, often custom, bioinformatic processing.

Computational Limitations

Perhaps the most pervasive barrier is the immense computational resource demand of WGS analysis. The volume of data generated is staggering; a single whole genome at 30x coverage can produce hundreds of gigabytes of raw data. Processing this data,through quality control, alignment, duplicate marking, and variant calling,requires substantial CPU power, large amounts of RAM, and vast storage space. For example, the alignment of hundreds of samples to a reference genome and subsequent joint variant calling are highly memory-intensive processes that typically require high-performance computing (HPC) clusters (Formenti et al., 2022). The accessibility of HPC resources is a major hurdle for many research communities focused on non-model species. While large genomic consortia have ready access to such infrastructure, individual research groups, particularly those in smaller institutions or developing countries, may not. This creates a significant disparity in the ability to conduct state-of-the-art genomic research. The computational burden also limits the exploration of different analytical parameters and the use of more accurate but computationally expensive software tools, potentially forcing researchers to adopt suboptimal methods. Furthermore, the expertise required to manage these computational workflows and large datasets presents a steep learning curve, creating a bioinformatic bottleneck that can slow the pace of discovery and innovation in non-model species genomics. Major bioinformatic challenges encountered in non-model species and commonly applied mitigation strategies are summarized in Figure 2.

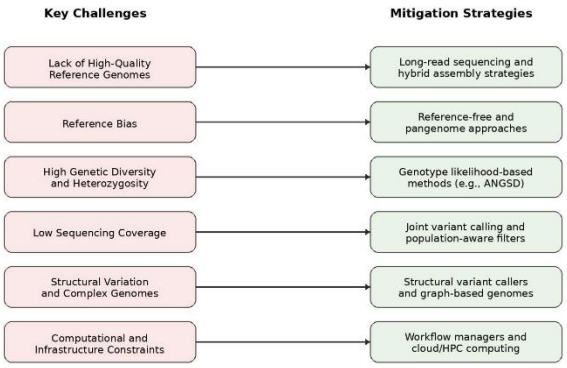

Figure 2: Key bioinformatic challenges and mitigation strategies in non-model genomics.

COMPONENTS OF BIOINFORMATIC PIPELINES FOR WHOLE GENOME SEQUENCING

The transformation of raw sequencing reads into biologically interpretable data requires a multi-stage bioinformatic pipeline. Each stage presents unique challenges and requires careful consideration when applied to non-model species. A robust pipeline is not merely a chain of tools but an integrated workflow where the output of each step critically influences the next. This section deconstructs the standard pipeline, examining the key components,from initial data preprocessing to final validation,and highlights the specific considerations and best practices for their application in non-model genomic studies.

Data Preprocessing

The initial and crucial step in any WGS analysis is data preprocessing, which aims to ensure that only high-quality data proceeds downstream, thereby reducing artifacts and improving the accuracy of all subsequent analyses. The process begins with quality control using tools like FastQC, which provides a visual report on read quality scores, nucleotide composition, adapter contamination, and the presence of over-represented sequences. For non-model species, particular attention must be paid to GC content, as significant deviations from the expected distribution can indicate contamination or specific technical biases. Following quality assessment, adapter trimming and filtering of low-quality reads and bases are performed using tools such as Trimmomatic or Cutadapt. This step is vital for removing sequencing adapters, which, if left in place, can align to the reference genome and create false positive variant calls, particularly indels. Furthermore, trimming low-quality bases from the ends of reads increases the accuracy of subsequent alignment. The stringency of filtering must be balanced; overly aggressive trimming can shorten reads to the point where they become unmappable, especially to a divergent reference genome. The parameters for these tools (e.g., sliding window quality thresholds, minimum read length) often require optimization based on the initial quality reports and the specific sequencing technology used (Bolger et al., 2014).

Read Alignment

The core step of mapping preprocessed reads to a reference genome is profoundly influenced by the nature of the non-model species' genome. The choice of aligner is critical. Burrows-Wheeler Aligner (BWA-MEM) and Bowtie2 are among the most widely used aligners due to their accuracy, efficiency, and sensitivity. BWA-MEM is generally preferred for its better performance with longer reads and indels, which are common when dealing with divergent genomes (Li, 2013). Handling an incomplete or divergent reference genome is the primary challenge. When using a reference from a closely related species, it is often necessary to adjust alignment parameters to allow for a higher mismatch rate. However, this increases the risk of reads from paralogous regions mapping incorrectly. A key strategy is to perform soft clipping, which allows the ends of reads that do not map to be excluded from the alignment without discarding the entire read, preserving information for variant calling. The resulting Sequence Alignment/Map (SAM) or its compressed counterpart (BAM) file must then be processed by marking PCR duplicates and, if possible, performing local realignment around indels. However, tools for these steps, like those in the Genome Analysis Toolkit (GATK) suite, are optimized for human data and may require significant parameter tuning for non-model organisms to avoid removing true biological variation (Poplin et al., 2018).

Variant Calling

Notably, no single pipeline performs optimally across all non-model systems, underscoring the necessity of context-dependent pipeline selection rather than reliance on default or human-centric workflows. Variant calling identifies genomic positions that differ from the reference genome, primarily focusing on single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels). This step is highly sensitive to alignment quality and read depth. Two main classes of variant callers are employed: those that operate on a single sample (e.g., BCFtools mpileup) and those designed for population-level variant detection by jointly calling variants across multiple samples (e.g., GATK HaplotypeCaller, FreeBayes). Joint calling is generally preferred as it improves sensitivity for detecting low-frequency variants by leveraging information across the entire cohort. The choice of tool involves trade-offs. GATK's HaplotypeCaller uses a powerful local assembly step that is excellent for calling indels and variants in complex regions but is computationally intensive and may be over-parameterized for non-human data. FreeBayes is a popular haplotype-based alternative that is often used in non-model species research due to its simpler model and fewer assumptions about ploidy and population structure (Garrison & Marth, 2012). For all callers, the resulting raw variant call format (VCF) file contains many false positives and must undergo rigorous filtering based on depth, quality scores, mapping quality, and strand bias. Establishing these thresholds without known truth sets is a major challenge and often relies on heuristic filters and visual inspection of the data. A comparative overview of commonly used variant calling pipelines and their suitability for non-model species is provided in Table 1.

Annotation and Functional Analysis

Determining the functional consequence of identified variants is a primary goal for many studies but is exceptionally difficult for non-model species. The challenge in gene annotation stems from the lack of well-annotated reference genomes. Without comprehensive databases of known genes and regulatory elements, predicting whether a variant is synonymous, non-synonymous, or in a regulatory region is fraught with uncertainty. Researchers must therefore rely on a combination of tools. Basic annotation involves mapping variant positions to any available gene predictions (GFF/GTF files) for the reference. For functional inference, tools like BLAST are used to find homologous sequences in model organism databases (e.g., UniProt,

RefSeq). InterProScan can then be used to predict protein domains and functional sites, while databases like eggNOG provide functional orthology assignments across a wide range of species (Huerta-Cepas et al., 2019). This comparative approach is powerful but imperfect; it risks misannotating genes that are novel to the species or have diverged in function, and it provides little insight into non-coding regulatory variants.

Table 1: Comparative overview of major variant calling pipelines and their suitability for non-model species.

Pipeline	Variant types supported	Key strengths	Key limitations	Computational demand	Recommended use cases in non-model / agricultural species
GATK	SNPs, small indels	High accuracy; extensive validation; strong community support	Optimized for human genomics; sensitive to reference bias	High	High-quality reference genomes; resequencing in well-annotated crops or livestock
FreeBayes	SNPs, indels	Flexible ploidy support; suitable for diverse genomes	Parameter-sensitive; limited structural variant detection	Moderate	Non-model species with variable ploidy and moderate coverage
SAMtools / BCFtools	SNPs, indels	Fast; lightweight; widely used	Lower sensitivity at low coverage; limited advanced filtering	Low	Exploratory analyses; preliminary variant discovery
ANGSD	Genotype likelihoods	Robust to low coverage; avoids hard genotype calls	Does not output genotypes directly	Low–Moderate	Population genomics; low-coverage agricultural datasets
dDocent	SNPs	Integrated pipeline; optimized for high heterozygosity	Complex setup; limited scalability	Moderate	Marine, wild relatives, and heterogeneous agricultural populations

Quality Control and Validation

Throughout the pipeline, rigorous quality control is essential. Metrics for assessing performance include alignment rates (the proportion of reads that map to the reference), mean depth of coverage, the distribution of depths, and the transition/transversion (Ti/Tv) ratio for SNP datasets,a ratio that typically falls within a predictable range for true variants and deviates for error-prone data. A low alignment rate or uneven coverage can indicate problems with the reference genome or library preparation. The most significant challenge is the validation of variants without a gold-standard dataset. Where resources allow, a subset of variants can be validated using an independent technology like Sanger sequencing. A powerful computational strategy is to use simulated datasets, where reads are generated in silico from a known genome sequence containing predefined variants. By running these simulated reads through the entire

pipeline, researchers can benchmark its sensitivity and false discovery rate and optimize parameters accordingly (Escalona et al., 2016). This process, though computationally expensive, is invaluable for developing a reliable and validated bioinformatic workflow tailored to the specific peculiarities of a non-model species study system.

Table 2: Comparative strengths, limitations, and recommended applications of major whole genome sequencing pipelines used in non-model and agricultural species.

Pipeline / framework	Typical inputs	Strengths	Limitations	Computational demand	Recommended applications (non-model / agricultural)
GATK Best Practices	BAM/CRAM aligned reads; reference genome; known sites optional	High accuracy; extensive documentation; broad community support	Human-centric defaults; requires high-quality reference; reference bias risk	High	Well-annotated crops/livestock; high-coverage resequencing; clinical-like diagnostics
ANGSD (genotype likelihood)	Aligned reads (BAM); reference recommended but flexible; low coverage	Robust to low coverage; avoids hard genotype calls; good for population inference	Requires statistical expertise; limited direct genotype output	Low–Moderate	Low-coverage population genomics; wild relatives; breeding populations under cost constraints
dDocent	Raw reads; reference optional; designed for high heterozygosity	Integrated workflow; optimized for diverse/non-model genomes	More complex configuration; limited scalability for very large datasets	Moderate	Highly heterozygous species; mixed breeding populations; non-model wildlife/agricultural systems
Snakemake / Nextflow-based custom workflows	Flexible (raw reads to variants/annotations); modular tool selection	Highly reproducible; scalable; portable across HPC/cloud; parameter transparency	Quality depends on design; requires pipeline development skills	Moderate–High (depends on tools)	Large multi-sample projects; institutional breeding programs; standardized reanalysis pipelines
Galaxy (web-based platform)	Raw reads or aligned reads; GUI-based analysis	Accessible for non-specialists; good for teaching and smaller projects	Limited scalability for very large WGS; depends on server resources	Low–Moderate	Small to medium datasets; capacity building in agricultural genomics labs
Reference-free assembly + variant discovery (hybrid)	Short + long reads; assembly graphs; optional reference anchoring	Reduces reference bias; captures SVs; improves genome representation	Computationally intensive; requires careful QC and validation	High	Species lacking references; crop wild relatives; structural variation-driven traits

EXISTING BIOINFORMATIC PIPELINES

Comparative Evaluation of Major Bioinformatic Pipelines

General-purpose pipelines such as GATK provide high accuracy and extensive validation but are computationally intensive and optimized primarily for human genomics. In contrast, pipelines such as ANGSD and dDocent prioritize robustness to low sequencing coverage and high heterozygosity, making them more suitable for non-model and agricultural species where reference quality and sequencing depth are often limited. The complexity of analyzing whole genome sequencing (WGS) data has led to the development of standardized bioinformatic pipelines. These frameworks aim to streamline the analytical process, reduce human error, and enhance reproducibility. For researchers working with non-model species, the choice of pipeline is critical and must be guided by an understanding of their underlying assumptions, strengths, and limitations. Existing solutions range from highly polished, general-purpose frameworks designed for human genetics to specialized tools built specifically to handle the challenges of diverse, poorly referenced genomes. Furthermore, the rise of workflow management systems has empowered researchers to construct robust, scalable, and reproducible analytical pathways, even for the most complex non-model projects. A comparative summary of major WGS pipeline frameworks and their suitability for non-model and agricultural species is presented in Table 2.

General-Purpose Pipelines

The gold standard in human genomics is the GATK Best Practices pipeline, developed by the Broad Institute. This comprehensive framework provides a meticulously validated series of steps for data preprocessing, alignment, base quality score recalibration (BQSR), and variant calling using the HaplotypeCaller in a joint-genotyping approach. Its rigorous methodology minimizes artifacts and produces exceptionally high-quality variant calls for human data. Similarly, the suite of tools within SAMtools and BCFtools, pioneered by Heng Li, offers a more modular but widely adopted set of utilities for processing alignments (samtools) and calling variants (bcftools mpileup). However, the direct applicability of these general-purpose pipelines to non-model species is limited. The GATK Best Practices workflow makes several key assumptions that are often violated in non-model systems. The BQSR step, for instance, requires a known database of polymorphic sites to recalibrate base quality scores,a resource that is absent for non-model organisms. Furthermore, the HaplotypeCaller's statistical models are finely tuned for human levels of heterozygosity and specific error profiles. When applied to a highly diverse non-model species or a divergent reference genome, these models can perform suboptimally, leading to a high false positive rate or an under-calling of true variants (Poplin et al., 2018). While these tools can often be used as components within a larger workflow, their "best practices" require significant modification and parameter tuning to be effective outside of the human context.

Specialized Pipelines for Non-Model Species

Recognizing the limitations of general-purpose tools, the community has developed several specialized pipelines explicitly designed for the challenges of non-model species. These tools often forego the need for a high-quality reference genome or are built to handle high levels of diversity and missing data. A seminal example in the realm of reduced-representation sequencing (RAD-seq) that has influenced WGS approaches is Stacks. While designed for restriction-site-associated DNA sequencing, its philosophy of reference-free locus discovery and genotyping without strict dependence on a reference genome has been foundational. For WGS, pipelines like dDocent have gained significant traction. dDocent is a flexible, open-source workflow that guides users from raw WGS reads to validated SNPs. Its strength lies in its adaptability; it can perform reference-free reference assembly from the data itself, align reads to this reference-free assembly or an existing reference, and call variants using a combination of FreeBayes and other tools (Puritz et al., 2014). It includes built-in filters for quality and balance of allele depths, which are crucial for managing high heterozygosity. For low-coverage WGS data or projects where genotype likelihoods are preferable to called genotypes due to uncertainty, ANGSD (Analysis of Next Generation Sequencing Data) is a powerful framework. ANGSD does not call genotypes explicitly. Instead, it calculates genotype probabilities and uses these likelihoods to estimate key population genetics parameters like allele frequencies, PCA, and admixture proportions directly. This approach is particularly valuable for non-model species as it is more robust to low coverage and avoids the biases introduced by hard genotype calling filters (Korneliussen et al., 2014). These specialized pipelines share a common feature: they prioritize flexibility and robustness to missing data and technical artifacts over the maximum possible precision achievable in ideal model organism settings.

Workflow Management Systems

Beyond pre-packaged pipelines, a modern approach involves building custom workflows using workflow management systems such as Snakemake and Nextflow. These systems allow researchers to encode their entire bioinformatic pipeline,from quality control to variant

calling,in a single, executable script. They manage the execution of each step, automatically handling software dependencies (often via containers like Docker or Singularity) and ensuring that if a run fails or new data is

Table 3: Commonly used bioinformatic tools for read alignment, variant calling, and annotation in non-model species, with key assumptions and limitations.

Pipeline stage	Tool	Primary function	Key assumptions	Major limitations	Typical applications in non-model / agricultural genomics
Quality control	FastQC / MultiQC	Assess raw read quality and sequencing artifacts	Quality metrics reflect downstream performance	Does not correct errors; diagnostic only	Initial assessment of WGS data from crops, livestock, pathogens
Read alignment	BWA-MEM	Align short reads to a reference genome	Reference genome adequately represents sample	Sensitive to reference bias; less effective for SVs	Resequencing of crops/livestock with available references
Read alignment	Bowtie2	Fast short-read alignment	Low divergence between reads and reference	Reduced accuracy for highly divergent genomes	Population-scale resequencing with moderate diversity
De novo / reference-free assembly	SPAdes / Flye	Assemble genomes from short or long reads	Sufficient coverage and read quality	High computational demand; fragmented assemblies	Genome reconstruction for poorly characterized species
Variant calling	FreeBayes	Detect SNPs and indels using Bayesian models	Reasonable coverage and ploidy specification	Parameter-sensitive; limited SV detection	Non-model species with variable ploidy
Variant calling	GATK HaplotypeCaller	Accurate SNP and indel calling	High-quality reference and calibration data	Human-centric defaults; high computational cost	Well-annotated agricultural species
Low-coverage inference	ANGSD	Estimate genotype likelihoods and population statistics	Population-level inference preferred over genotypes	No direct genotype calls	Low-coverage population genomics in crops and wild relatives
Annotation	SnpEff / VEP	Predict functional effects of variants	Accurate gene models available	Poor performance with incomplete annotations	Functional interpretation in crops/livestock
Functional annotation	InterProScan / eggNOG	Assign protein domains and functional categories	Homology reflects function	Computationally intensive; incomplete databases	Trait-associated gene discovery

added, only the necessary steps are re-run. The benefits of reproducibility and scalability are profound. A Snakemake or Nextflow script acts as a complete and unambiguous record of the entire analysis, detailing every software version, parameter, and command used. This makes the analysis perfectly reproducible, a critical but often elusive standard in scientific computing. Furthermore, these workflows are designed for scalability. They can seamlessly execute on a single laptop, a high-performance computing cluster, or in the cloud, automatically managing job scheduling and parallelization without the researcher having to rewrite the pipeline for each environment (Mölder et al., 2021). This is invaluable for WGS projects involving dozens or hundreds of samples, where computational management becomes a major task. Platforms like Galaxy offer a complementary approach, providing a user-friendly, web-based interface for hundreds of bioinformatic tools. Galaxy is excellent for beginners or for prototyping analyses, as it removes the command-line barrier and tracks the history of all operations. However, for large-scale WGS projects, the scalability and granular control offered by Snakemake and Nextflow often make them the preferred choice for production-level analyses.

BEST PRACTICES AND OPTIMIZATION

Outstanding Challenges and Research Gaps

Despite substantial methodological advances, several research gaps remain unresolved in non-model genomics. These include the absence of standardized benchmarking datasets, limited empirical validation of variant calling accuracy across diverse taxa, and insufficient integration of genomic pipelines with phenotypic and agronomic data. Addressing these gaps is essential for improving reproducibility and practical utility. The construction and execution of a bioinformatic pipeline for non-model species is not a one-size-fits-all endeavor. It is an iterative process of design, optimization, and validation that must be tailored to the specific biological context and computational constraints of the study. Moving beyond the mere selection of tools, this phase involves strategic decisions about the pipeline's architecture, its computational footprint, and its capacity to generate biologically holistic insights. Adhering to best practices in pipeline design, computational efficiency, and data integration is paramount for ensuring that the analysis is robust, scalable, and ultimately capable of answering the complex questions posed by non-model organism genomics. The conceptual progression from current limitations to unresolved research gaps and future research directions in non-model genomics is illustrated in Figure 3. Commonly used bioinformatic tools across different stages of whole genome sequencing pipelines, along with their underlying assumptions and limitations, are summarized in Table 3.

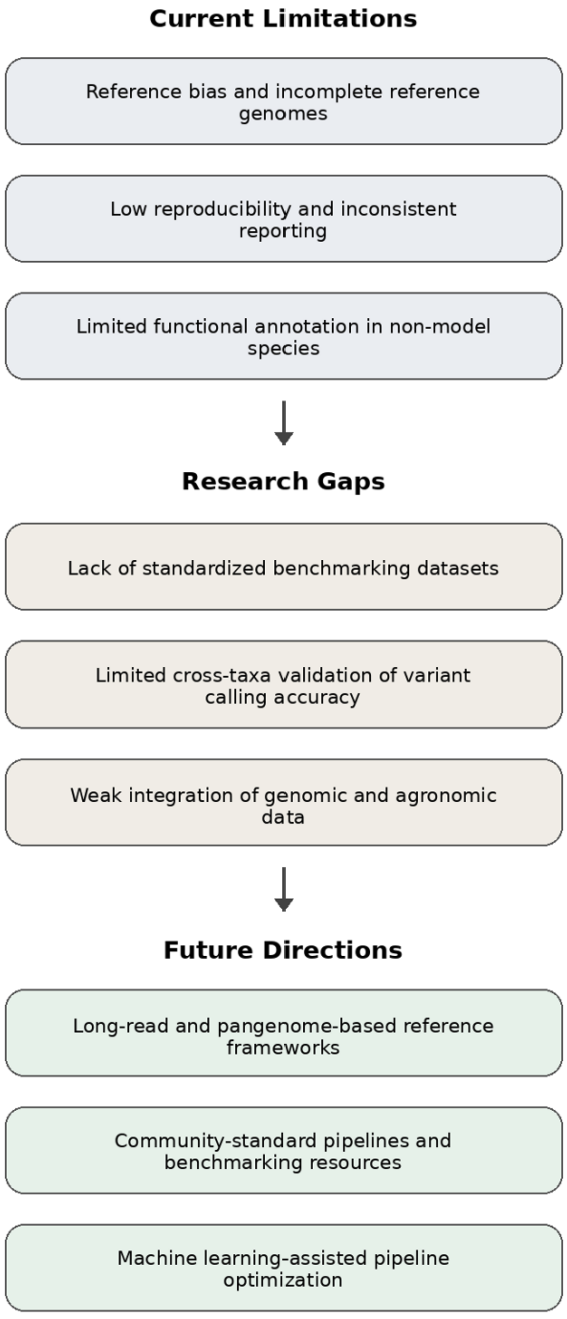

Figure 3: Conceptual framework summarizing current limitations, methodological gaps, and future research directions in whole genome sequencing bioinformatics for non-model species.

Pipeline Design Considerations

The foundational principle of effective pipeline design is modularity and flexibility. A modular pipeline is constructed as a series of independent, interchangeable components (e.g., quality control, alignment, variant calling) rather than a single, monolithic script. This architecture, often facilitated by workflow managers like Snakemake or Nextflow, allows researchers to easily swap tools or update specific steps without overhauling the entire workflow. For instance, one might test both BWA-MEM and Bowtie2 as aligners or compare GATK with FreeBayes for variant calling on a subset of data to determine the best performer for their specific genome (Mölder et al., 2021). This flexibility is essential for diverse datasets, as the optimal tool for a highly heterozygous invertebrate genome may differ from that for a inbred vertebrate population. Closely tied to modularity is the critical need for parameter optimization. Off-the-shelf software defaults are invariably tuned for human data and perform poorly on divergent non-model genomes. A systematic approach to optimization is required. This begins with generating a small, "truth set" for validation, which could involve Sanger sequencing of a few genomic regions, using simulated reads with known variants, or even leveraging high-quality data from a subset of samples. By running the pipeline with different parameters (e.g., mapping stringency, variant quality thresholds) and comparing the output to the truth set, researchers can empirically determine the settings that maximize the F1 score,the harmonic mean of precision (minimizing false positives) and recall (minimizing false negatives) (O'Neill et al., 2022). This process, while computationally demanding initially, is a non-negotiable best practice for ensuring data quality and is far superior to adopting default parameters or those from unrelated studies.

Computational Efficiency

The scale of WGS data makes computational efficiency a primary concern, especially for research groups without access to massive computing infrastructures. Fortunately, numerous strategies for reducing runtime and memory usage can be employed. A fundamental first step is pre-processing: rigorous quality trimming and filtering can drastically reduce the volume of data entering the alignment stage, saving substantial time and storage. Choosing the right file formats is also crucial; converting SAM files to compressed BAM/CRAM formats reduces storage needs, and indexing these files enables rapid access. The most powerful strategy is parallelization. Most pipeline steps are "embarrassingly parallel," meaning individual samples or chromosomes can be processed simultaneously without interdependency. Workflow managers like Snakemake and Nextflow excel at automatically managing this parallelization, splitting jobs across multiple CPU cores on a cluster. For particularly demanding steps like sequence alignment or variant calling, selecting tools that are themselves multi-threaded (e.g., using the -t flag in BWA) can yield significant speed improvements. When local computational resources are saturated, cloud computing solutions (e.g., Amazon Web Services, Google Cloud Platform, Microsoft Azure) offer a powerful alternative. The cloud provides virtually unlimited, on-demand computing power, allowing researchers to scale their analysis to hundreds of samples by simply launching more instances. The key to cost-effective cloud usage is to choose instance types that match the task (e.g., high-memory instances for assembly, compute-optimized instances for alignment) and to use spot instances for fault-tolerant jobs to reduce costs by up to 90%. While cloud computing introduces complexities in data transfer and cost management, its flexibility is unmatched and is democratizing access to high-performance computing for non-model species research (Reid & Lapp, 2020).

Data Integration

The true power of modern genomics is realized not in isolation, but through integrative analysis. Combining WGS data with other omics data types provides a systems-level view of biological function that any single approach cannot achieve. For example, overlaying genome-wide SNPs with transcriptomics (RNA-seq) data from the same individuals can identify expression Quantitative Trait Loci (eQTLs), revealing genetic variants that regulate gene expression and providing mechanistic insight into putative adaptive loci identified in a GWAS. Similarly, integrating WGS with epigenomics data (e.g., ATAC-seq or bisulfite sequencing for DNA methylation) can uncover the regulatory landscape and show how genetic variation influences chromatin accessibility and epigenetic marks, which in turn affect phenotype (Hoffman & Williams, 2019). The challenge lies in the tools for integrative analysis, as non-model species lack the curated databases that facilitate this in models like human or mouse. The process often requires a bespoke bioinformatic approach. Key strategies include:

- Comparative Genomics: Using tools like BLAST, OrthoFinder, or Ensembl Compara to find orthologous genes and regulatory regions in model species, allowing for the transfer of functional annotation.
- Multi-Omics Alignment: Ensuring all data types (WGS, RNA-seq, etc.) are aligned to the same reference genome assembly to guarantee coordinate consistency.

- Custom Scripting: Writing scripts in R or Python to intersect variant calls (VCF files) with gene expression tables (from RNA-seq) or chromatin peak calls (from ATAC-seq) to find correlations and overlaps.

While complex, this integrated approach is the future of non-model species genomics. It moves beyond cataloging genetic variation to understanding its functional consequences, enabling researchers to connect genotype to phenotype through intermediate molecular layers and build a more comprehensive model of adaptation, response, and function in the organisms they study.

CHALLENGES AND LIMITATIONS

Rather than serving as a descriptive inventory of tools, this review emphasizes informed methodological decision-making, recognizing that pipeline performance is inherently dependent on species biology, data quality, and clearly defined research objectives. Within agricultural genomics, best-practice pipeline optimization directly influences the detection of quantitative trait loci, genomic estimated breeding values, and disease-associated variants. Ethical considerations are particularly important for indigenous crop varieties, local livestock breeds, and regionally adapted germplasm, where genomic data sharing must balance open science principles with data sovereignty and community rights. Reproducibility should be treated as a baseline requirement rather than a best-case outcome, necessitating transparent reporting of software versions, parameter settings, reference assemblies, and variant filtering criteria.

Despite the remarkable advances in sequencing technologies and bioinformatic tools, the analysis of whole genome sequencing (WGS) data for non-model species remains fraught with significant challenges that extend beyond mere technical execution. These limitations often reside in the human, ethical, and logistical dimensions of research, presenting barriers that can hinder progress and equitable participation in the genomic revolution. Addressing these issues is as critical as developing new algorithms, for they determine who can generate knowledge and how reliably it can be built upon.

Bioinformatic Expertise

A primary bottleneck in non-model species genomics is the acute need for training in pipeline development and interpretation. The field demands a rare hybrid of skills: deep biological knowledge of the study system coupled with computational proficiency in software engineering, statistics, and data management. Few academic programs adequately train biologists in these computational skills, creating a significant expertise gap. Researchers often find themselves spending more time debugging code, managing software dependencies, and configuring high-performance computing clusters than interpreting biological results (Leprevost et al., 2017). This steep learning curve can lead to the implementation of suboptimal methods or the misinterpretation of output data, potentially compromising the validity of scientific conclusions. This expertise barrier directly impacts the accessibility for researchers in resource-limited settings. The genomics of global biodiversity is often studied in countries with rich biodiversity but limited computational infrastructure and funding. These researchers face a double burden: the high cost of sequencing and the even greater challenge of analyzing the data without access to bioinformaticians, high-performance computing, or stable internet connections. This creates a concerning disparity where the species most in need of genomic research,those in threatened ecosystems,are often studied by teams from wealthier nations, potentially perpetuating a form of "scientific colonialism" where data is extracted without building local capacity (Hogg et al., 2022). Bridging this gap requires intentional efforts in training, resource sharing, and the development of less computationally intensive methods.

Standardization and Reproducibility

The field of non-model genomics suffers from a profound lack of standardization, leading to high variability in pipeline outputs across studies. The same raw dataset processed through different pipelines,or even the same pipeline with different parameters,can yield vastly different variant sets and subsequent biological inferences. For instance, the choice of mapping stringency, variant caller, and quality filters can alter estimates of population genetic parameters like nucleotide diversity (π) and Tajima's D, which are central to testing evolutionary hypotheses (O'Neill et al., 2022). This lack of consistency makes it difficult to compare results across studies or perform meaningful meta-analyses, fragmenting the field and slowing cumulative progress. Consequently, the importance of standardized reporting and documentation cannot be overstated. Reproducibility, a cornerstone of the scientific method, is exceptionally difficult to achieve in computational biology. It requires not just sharing code but comprehensively documenting every software version, parameter setting, and reference genome used. Best practices now advocate for the use of workflow managers (Snakemake, Nextflow) and containerization technologies (Docker, Singularity) that encapsulate the entire computational environment, ensuring that an analysis can be run identically years later (Mölder et al., 2021). Furthermore, adhering to reporting standards, such as those proposed for bioinformatic workflows, is essential for allowing others to understand, evaluate, and build upon published work. Key reporting elements required to ensure transparency and reproducibility of whole genome

sequencing bioinformatic analyses are summarized in Table 4.

Table 4: Reporting and reproducibility checklist for whole genome sequencing bioinformatic pipelines applied to non-model and agricultural species

Category	Item to report	Why it matters	Recommended reporting practice
Sequencing data	Platform, read length, coverage, library preparation	Affects error profiles, variant detection, and reproducibility	Report in Methods with accession numbers where applicable
Reference genome	Assembly version, source, annotation status	Influences alignment accuracy and reference bias	Specify reference build and justification for selection
Quality control	Filtering thresholds and QC tools used	Ensures transparency in data exclusion and preprocessing	Provide exact parameters and summary statistics
Alignment / assembly	Software, version, and parameters	Strongly impacts downstream variant calling	List tools, versions, and non-default parameters
Variant calling	Caller, model assumptions, ploidy settings	Determines sensitivity and specificity of detected variants	Report caller choice and rationale
Filtering criteria	Hard filters or statistical thresholds	Affects false positive and false negative rates	Provide thresholds and justification
Annotation	Databases and annotation tools used	Determines functional interpretation of variants	Specify database versions and annotation pipelines
Workflow management	Use of Snakemake, Nextflow, or equivalent	Improves reproducibility and scalability	Describe workflow framework and execution environment
Computational environment	Hardware, OS, containerization	Ensures analyses can be reproduced	Report HPC/cloud resources and container images
Data and code availability	Repositories for raw data, scripts, workflows	Supports transparency and reuse	Provide persistent links (e.g., ENA, NCBI, GitHub)

Ethical and Data Sharing Issues

The generation of genomic data from non-model species, particularly those that are endangered or culturally significant, introduces complex ethical considerations. Publishing the full genome of an endangered species could theoretically provide a blueprint for its exploitation (e.g., by revealing genes for valuable traits) or could facilitate biopiracy. There is an ongoing debate within the conservation genomics community about how to balance the imperative of open science with the need to protect vulnerable species. Practices such as depositing data in managed-access repositories (e.g., NCBI's dbGaP) or releasing only a masked version of the genome are emerging as potential solutions, though they challenge the norm of full data openness (Formenti et al., 2022). Despite these concerns, promoting open-access tools and repositories remains a fundamental principle for advancing the field. The development of bioinformatic software as open-source projects allows for community scrutiny, improvement, and adaptation. Similarly, archiving data in public repositories like the NCBI Sequence Read Archive (SRA) and GenBank is crucial for preventing data loss, enabling reproducibility, and allowing the global research community to extract maximum value from expensive sequencing projects. The challenge is to develop nuanced data-sharing policies that respect sovereignty and conservation concerns while upholding the ethos of collaborative, open science.

FUTURE DIRECTIONS

Future research priorities include the development of standardized benchmarking datasets, broader representation

of agriculturally relevant species in genomic databases, and the integration of machine learning approaches for automated pipeline optimization. The future of bioinformatic pipelines for non-model species is bright,

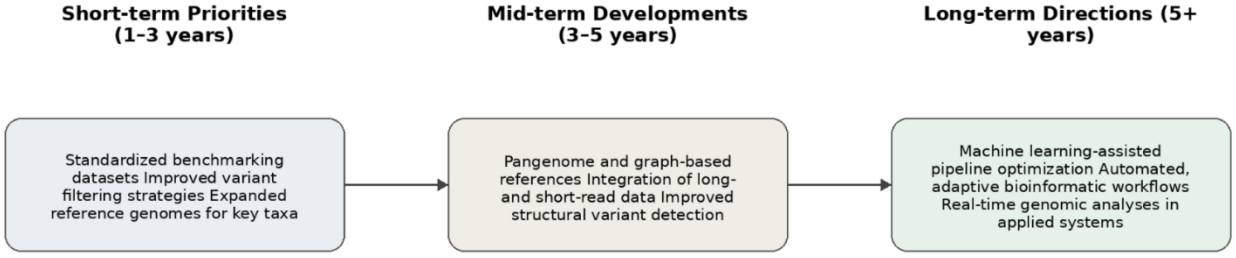

Figure 4: Future research priorities and emerging methodological trends in whole genome sequencing bioinformatics for non-model and agricultural species.

shaped by rapid technological innovation and a growing awareness of the need for collaboration and accessibility. Several key trends are poised to address current limitations and open new frontiers of discovery. The continued maturation of long-read sequencing technologies from PacBio and Oxford Nanopore is a game-changer. These technologies produce reads that are thousands to millions of bases long, effortlessly spanning repetitive regions and complex structural variations that confound short-read assemblers. The impact on non-model species is profound: it is now feasible for individual labs to generate high-quality, chromosome-level genome assemblies without the need for expensive ancillary techniques like Hi-C, providing a robust foundation for all downstream resequencing analyses (Hotaling et al., 2021). The integration of these long reads into WGS pipelines will improve mapping fidelity and variant calling accuracy, particularly for indels and SVs, finally allowing researchers to fully characterize the pan-genome of diverse species populations. Machine learning (ML) is set to revolutionize many aspects of the bioinformatic pipeline. Supervised learning models can be trained to distinguish true genetic variants from sequencing artifacts with higher accuracy than traditional statistical filters, leading to cleaner variant calls. In genome annotation, deep learning models like AlphaFold2 for protein structure prediction are being adapted to predict functional elements and the regulatory impact of non-coding variants, even in the absence of experimental data for that species (Jumper et al., 2021). Perhaps most intriguingly, ML holds the potential for automated pipeline optimization, where algorithms could intelligently test thousands of parameter combinations to identify the optimal workflow for a given dataset, removing a major source of subjectivity and manual effort. Future research priorities and emerging methodological trends in whole genome sequencing bioinformatics for non-model species are summarized in Figure 4.

Addressing the challenges of non-model species requires a collective effort. Community-driven initiatives are increasingly powerful. Consortia such as the Vertebrate Genomes Project (VGP) and the Earth BioGenome Project (EBP) are establishing standardized, high-quality pipelines for genome assembly and annotation that will serve as benchmarks for the entire field. Open-source platforms like GitHub and BioConda are essential for sharing code and managing software distributions, respectively. A critical future direction is the concerted effort toward expanding genomic databases for non-model species, creating centralized resources that aggregate genomes, variants, and functional annotations, making comparative analyses far more efficient and powerful. To democratize access, there is a growing push to develop portable and user-friendly tools. This includes the development of GUI-based pipelines for non-experts, which hide the underlying command-line complexity behind intuitive graphical interfaces. Platforms

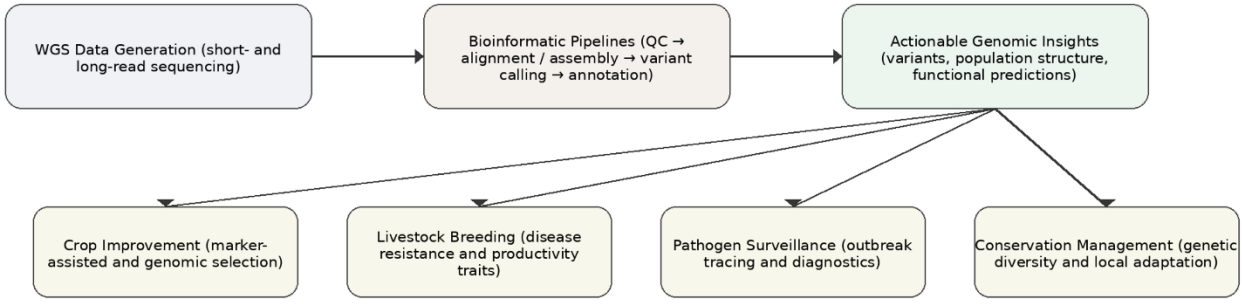

Figure 5: Translational pathways linking whole genome sequencing bioinformatic pipelines to applied agricultural and conservation workflows.

like Galaxy already offer this, but future tools will need to be even more specialized and well-documented for specific non-model applications. Furthermore, cloud-based solutions for global access are eliminating the need for expensive local hardware. Cloud platforms can offer pre-configured, scalable virtual machines with popular pipelines already installed, allowing researchers anywhere with an internet connection to analyze large genomic datasets by paying only for the computing time they use. The translational pathways linking whole genome sequencing bioinformatic pipelines to applied agricultural and conservation workflows are illustrated in Figure 5.

CONCLUSION

Whole genome sequencing has transformed the study of genetic variation in non-model species, offering unprecedented opportunities for advancing evolutionary research, conservation efforts, and agricultural improvement. However, the effectiveness of WGS in these systems depends not only on sequencing technologies but also on the careful design and implementation of bioinformatic pipelines that account for biological complexity, data limitations, and analytical trade-offs. As this review has demonstrated, no single pipeline or tool is universally optimal for all non-model species, underscoring the necessity of context-dependent methodological choices informed by species biology, study objectives, and resource availability.

Persistent challenges, including reference bias, high heterozygosity, low or uneven sequencing coverage, and limited functional annotation, continue to shape the accuracy and interpretability of genomic analyses. Addressing these issues requires greater emphasis on transparent reporting, standardized benchmarking, and reproducible workflow design. In agricultural genomics, where WGS data increasingly inform crop improvement, livestock breeding, and pathogen surveillance, these considerations are particularly critical, as analytical biases can directly influence breeding decisions and management strategies.

Looking forward, the integration of long-read sequencing, pangenome representations, and scalable workflow management systems is expected to reduce reference dependence and improve variant detection across diverse taxa. Emerging machine learning approaches and community-driven benchmarking initiatives further offer promising avenues for improving pipeline robustness and automation. Ultimately, progress in non-model genomics will depend on sustained efforts to align bioinformatic innovation with biological realism and applied needs. By synthesizing current practices, limitations, and future directions, this review provides a framework for advancing reliable and impactful whole genome sequencing analyses in non-model and agricultural species.

DECLARATIONS

AI Usage Declaration

In line with COPE guidelines, AI-assisted tools were used only for language editing and formatting and did not contribute to scientific content, data, analysis, or conclusions. All responsibility for the manuscript rests with the authors.

REFERENCES

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics, *17*(8), 459–469. https://doi.org/10.1038/nrg.2016.57

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Sadye, P., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., ... & European Reference Genome Atlas (ERGA) Consortium. (2022). The era of reference genomes in conservation genomics. Trends in Ecology & Evolution, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.005

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint. arXiv:1207.3907. https://arxiv.org/abs/1207.3907

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLOS Genetics, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. Molecular Ecology, *30*(1), 62–82. https://doi.org/10.1111/mec.15720

Hoffman, J. I., & Williams, C. L. (2019). A framework for integrating multiple omics datasets to identify genomic features that predict disease risk. Briefings in Bioinformatics, *20*(4), 1301–1312. https://doi.org/10.1093/bib/bbx173

Hogg, C. J., Ottewell, K., Latch, P., Rossetto, M., Biggs, J., Gilbert, A., Godwin, J., Gross, J., Hoeben, P., Holleley, C. E., Hunter, D. A., Lacy, R. C., Lott, M. J., Mastrantonis, S., McDonald, P. G., McLennan, E. A., Peel, E., Pellatt, E. J., Percival-Alwyn, L., ... & Grueber, C. E. (2022). Genomics for conserving threatened species: bridging the gap between theory and practice. Nature Reviews Genetics, *23*(6), 381–393. https://doi.org/10.1038/s41576-022-00458-7

Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are this section now? Proceedings of the National Academy of Sciences, *118*(52), e2109019118. https://doi.org/10.1073/pnas.2109019118

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold2. Nature, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Leprevost, F. da V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., & Carvalho, P. C. (2017). BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics, *33*(16), 2580–2582. https://doi.org/10.1093/bioinformatics/btx192

Leroy, G., Carroll, E. L., Bruford, M. W., DeWoody, J. A., Strand, A., Waits, L., & Wang, J. (2021). Next-generation metrics for monitoring genetic erosion within populations of conservation concern. Evolutionary Applications, *14*(5), 1238–1245. https://doi.org/10.1111/eva.13190

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint. arXiv:1303.3997. https://arxiv.org/abs/1303.3997

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology, *30*(23), 5966–5993. https://doi.org/10.1111/mec.16077

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. Genome Biology, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. F1000Research, *10*(33). https://doi.org/10.12688/f1000research.29032.2

O'Neill, M. J., Lawton, B. R., & Rehan, S. M. (2022). Biased representation of genetic variation in non-model species: An evaluation of SNP panels for conservation genomics. Conservation Genetics, *23*(2), 247–260. https://doi.org/10.1007/s10592-021-01415-5

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv, 201178. https://doi.org/10.1101/201178

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ, *2*, e431. https://doi.org/10.7717/peerj.431

Reid, J. G., & Lapp, H. (2020). Bioinformatic strategies for analyzing ultra-large-scale sequence data. Current Protocols in Bioinformatics, *70*(1), e102. https://doi.org/10.1002/cpbi.102

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoeppner, M. P., Höglund, J., Jensen, E. L., Krause, J., Kristensen, T. N., ... & Zieliński, P. (2017). Genomics and the challenging translation into conservation practice. Trends in Ecology & Evolution, *32*(2), 81–92. https://doi.org/10.1016/j.tree.2016.11.006