

STANDARDS RESEARCH

Agricultural Genomics Data: Assessing the Opportunity for Data Standards

May 2026

Authors

John Crowley, Ph.D., AbacusBio Ltd.

Doug Bjelland, Ph.D., AbacusBio Ltd.

Fiona Hely, BAppSc. AbacusBio Ltd.

Tiago Valente, Ph.D., AbacusBio Ltd.

Marcelo Mollinari, Ph.D., AbacusBio Ltd.

Project Advisory Panel

Loreta Gudynaite, Bayer CropScience Inc.

Michelle Miller, Independent Consultant

Paul Stothard, University of Alberta

Rocío Morales Rayas, CSA Group

Samuel Sulemanu, CSA Group (Project Manager)

Acknowledgements

The contributions of Catherine Tays, Marc Karam and Natalia Ortiz from CSA Group during the initial stages of this project are acknowledged.

Financial Support

This work was supported by the Government of Ontario through the Ontario Research Fund and the Government of Canada through Genome Canada.

Disclaimer

This work has been produced by the authors and is owned by the Canadian Standards Association. It is designed to provide general information in regard to the subject matter covered. The views expressed in this publication are those of the authors and study participants. The authors and Canadian Standards Association are not responsible for any loss or damage which might occur as a result of your reliance or use of the content in this publication.

Table of Contents

Executive Summary 5

1. Introduction 7

2. Methods 9

2.1. Literature Review and Landscape Analysis 9

2.2. Insights from International Interviews 9

2.3. Interdisciplinary Challenge Teams 10

2.3.1. Information Collection Matrix 10

2.4. Workshop and Validation 11

3. Results and Discussion 12

3.1. Literature Review and Landscape Analysis 12

3.1.1. Best Practices and Processes for Genomic Data Collection 12

3.1.2. Best Practices and Processes for Genomic Data Management, Reporting, and Sharing 15

3.1.3. Best Practices and Processes for Data Analysis 19

3.1.4. Best Practices and Processes for Data Governance 21

3.2. International Projects 23

3.2.1. International Interviewees 23

3.2.2. Interviews and Feedback 23

3.3. Genomic Technologies and Data from the Interdisciplinary Challenge Teams 26

3.3.1. Interdisciplinary Challenge Teams 26

3.3.2. Potential Benefits of Interdisciplinary Challenge Team Outcomes 29

3.3.3. Potential End Users of Interdisciplinary Challenge Team Outcomes 31

3.3.4. Interdisciplinary Challenge Team Information Collection Matrix 32

3.3.5. Interdisciplinary Challenge Team Data Collection, Management and Sharing 32

3.3.6. Interdisciplinary Challenge Team Governance and Metadata 35

3.3.7. Interdisciplinary Challenge Team National and International Uptake 36

3.3.8. Existing and Future Challenges for Interdisciplinary Challenge Team Data 39

3.4. Use Cases for Standardized Interdisciplinary Challenge Team Data 41
3.5. Roadblocks and Interdisciplinary Challenge Team Data Gaps 42
3.6. Recommendations for Interdisciplinary Challenge Team Data Tools 43
3.6.1. Implement Cross-Project Governance and Coordination 44
3.6.2. Develop and Adopt Unified Metadata Templates 44
3.6.3. Harmonize Data Processing and Quality Control 45
3.6.4. Create a Shared Discovery Catalogue and Repository Policy 46
3.6.5. Enable Cross-Project Data Integration and Reuse 46
3.6.6. Support Training and Capacity Building 47
3.6.7. Monitor, Evaluate, and Refine Implementation 47
3.6.8. Secure Funding for Long-Term Data Preservation 48
3.7 Workshop Validation and Interviews 49
3.7.1. Use of Metadata Templates 53
3.7.2. Cross-Project Data Governance 54
3.7.3. Data Ownership and Sovereignty 55
3.7.4. Refine Standardization Needs 56
4. Recommendations for Agricultural Genomics and Conclusions 57
References 61
Appendix A—Questionnaire for the International Interviews 70
Appendix B—Questionnaire for the Interdisciplinary Challenge Teams 72
Appendix C—Public Data Management Example in the United States: Ag Data Commons 75
Appendix D—Questionnaire for the Workshop Follow-Up Interviews 77

Executive Summary

Modern agriculture can benefit from the use of genomic technologies to enhance efficiency, productivity, and health. Macro goals include those that address challenges related to environmental impacts and climate change. While genomic technologies generate large, often complex sets of genomic, phenotypic, and demographic data or metadata, their full potential value can be limited by inconsistent data practices, fragmented data formats, and unclear governance. This research report examines the evolving landscape of agricultural genomics (agri-genomics) data practices to assess the opportunity for data standards, focusing on genomic data governance, management, storage, and sharing.

The research conducted for this report included a comprehensive literature review and landscape analysis, interviews with international experts, screening of data practices implemented by Genome Canada-funded Interdisciplinary Challenge Teams, a workshop, and workshop follow-up interviews. This report identifies both challenges and opportunities for harmonizing genomic data and associated practices, which are essential for implementing standards in Canada. The key barriers addressed in this report include a lack of or inconsistent metadata templates and descriptors, unclear data governance and ownership, insufficient training opportunities for data producers and users, and the need to secure funding for long-term data preservation. In addition, the adoption of FAIR (findable, accessible, interoperable, reusable) principles is emphasized as essential to maximizing the impact of genomic resources and fostering collaboration among researchers, producers, and service providers.

This report also discusses and highlights the importance of standardizing data practices to enable effective cross-project data reuse, data longevity, interoperability, and equitable access, particularly for publicly-funded initiatives. Finally, this report presents the following eight actionable recommendations for data standardization in agri-genomics:

1. Establish a cross-project governance committee that will clarify data governance frameworks and ownership guidelines to promote transparency, responsible use, and compliance with ethical and legal standards.
2. Create and adopt unified metadata templates and descriptors to ensure consistency and to facilitate data sharing across projects, organizations, and sectors.
3. Establish standardized protocols for data processing and quality control, so data quality and preparation steps are transparent even when analytical methods differ.
4. Create and adopt a shared discovery catalogue and repository policy, which will encourage equitable access to genomic data and innovations, especially for publicly-funded initiatives.
5. Facilitate cross-project data integration and reuse by adopting minimum integration specifications, consistent naming conventions, shared genome reference versions, practical integration guidelines, example pipelines, and other tools.

6. Develop and implement training programs and curricula for data generators and users to build capacity and to ensure best practices, from collection through to reporting.
7. Monitor, evaluate, and refine implementation of data standards using simple metrics and regular review cycles to ensure standards remain relevant as technologies evolve; explore international best practices while tailoring approaches to the Canadian context.
8. Secure sustainable funding for the long-term preservation, usability, and temporal interoperability of genomic datasets, enabling ongoing accessibility and reuse.

These recommendations, which can support responsible data use, enhance collaboration, and further initial and ongoing investments across the agricultural value chain, are tailored to the Canadian context, and provide a roadmap for data generators and users to unlock the full potential of genomics in the agriculture sector.



"Genomic studies play an important role in modern agriculture and have helped to increase the production and resilience of major crops and to improve the health and performance of poultry and livestock."

1. Introduction

Modern agriculture faces the challenge of adapting to climate change while maintaining high productivity and minimal environmental impact. Genomic technologies can help address this challenge by enabling a deeper understanding of agricultural systems [1]. However, because these technologies can produce large amounts of data, a shared language is required to ensure this valuable information can be consistently used across different projects within different disciplines and sectors. Genomics is characterized as the study of whole genomes as systems, including how genes are arranged and expressed across individuals. It also investigates how genes interact with one another and with the environment, and how these features can be mapped along chromosomes [2], [3]. Genomic studies play an important role in modern agriculture and have helped to increase the production and resilience of major crops and to improve the health and performance of poultry and livestock [4], [5]. Among those studies, the use of biological markers associated with traits, such as tolerance to biotic and abiotic stressors, production, and overall quality, has enabled the development of more efficient selection strategies both in plants and animals [5], [6], [7].

Agricultural genomics is the application of genomic technologies across crops, livestock, and agri-environment systems. In this field, a broad range of methodologies and applications are available, allowing biological samples to be obtained from single organisms or from multiple sources simultaneously, including individual specimens and environmental samples [4], [8]. Genomic data from crops, soil, microbes, poultry, and livestock are classified by the type of biomolecules they represent, including DNA, RNA, proteins, and small-molecule metabolites [9], [10]. In addition, genomic datasets can be classified by the technology used to generate them. For example, RNA can be quantified using microarrays or sequence-based assays, which produce gene expression data, or in some cases, gene regulation data [11]. As genomic technologies become more accessible, they enable broader generation and interpretation of large-scale datasets, expanding our understanding of biology and supporting trait discovery. In agriculture, these technologies commonly include sequencing-based genotyping, gene expression assays, and other high-throughput molecular profiling methods. These technologies also help translate research findings into more productive, sustainable, and economically relevant options for producers, service providers, and policy partners, who use genomics in their decision-making within the agri-food sector. Genomic technologies, along with supporting metadata (i.e., information that helps to describe, locate, manage, and understand collected research data), can help researchers develop crop varieties and livestock tailored to specific environments and market demands [5], [12], and are central to addressing climate change challenges, resource scarcity, and pathogen outbreaks [13]. While genomic

data are usually obtained to support specific goals, these valuable biological samples can be used for purposes beyond their original intent [14].

In Canada, Genome Canada's Climate-Smart Agriculture and Food Systems initiative has catalyzed interdisciplinary efforts to apply genomics to climate resilience and sustainability. In this report, the Interdisciplinary Challenge Teams (ICTs) funded by Genome Canada are featured as examples of how genomics data generated within specific research groups and fields can be used across ICTs and can provide value to public and private breeding programs, service provider laboratories, and tools used by producers. As part of this initiative, integrated teams of researchers are working toward a common goal to: "provide genomic solutions to help implement agriculture and food production systems that are climate-resilient, socially responsible, economically viable and environmentally sustainable, and that contribute to the mitigation of climate change impacts" [15, p. 1].

Even when genomic datasets are successfully produced and used in a specific project, their long-term preservation and data reuse is often hindered by inconsistent metadata, insufficient documentation, fragmented formats and pipelines, and unclear governance, among other factors. To facilitate cross-project reuse of genomic data generated by any research or commercial entity, and to encourage broader adoption of genomic technologies, there is a need for clear ethical and regulatory frameworks that address data privacy, consent, governance and management, and intellectual property concerns. Importantly, access to genomic data, results, and innovations must be equitable, especially in publicly-funded projects [16], [17]. This includes ensuring that producers, small plant and animal breeding programs, and service providers can access the data generated by research projects, even if they lack extensive internal informatics capacity. To minimize these obstacles, Wilkinson *et al.* [18] proposed the four foundational FAIR (findable, accessible, interoperable, reusable) principles, which serve as a guideline for data producers and publishers to improve the findability, accessibility, interoperability, and reusability of digital assets. These principles emphasize machine-actionability (i.e., the ability of computational systems to find, access, interoperate with, and reuse data with minimal human intervention) because there will be an increasing need to rely on computational support to deal with genomic data as its volume, complexity, and rate of creation increases. While FAIR principles improve technical aspects of reusing datasets, equitable access also depends on governance, permissions, and the capacity to reuse the data.

Despite the clear advantages of reusing data across projects, the lack of harmonized guidelines for genomic data collection, storage, and use in agriculture poses challenges. While some countries and organizations have developed data management and data sharing requirements and guidance (e.g., in Canada: [19], [20], [21], in USA: [22], the absence of unified protocols makes it difficult for researchers and producers to collaborate and share private and research data effectively within and across borders. This inconsistency can hinder the advancement of genomic research, limit the potential benefits for global food systems, and duplicate funding requirements where suitable data have already been collected. Thus, to fully leverage the power of genomics in agriculture, it is essential to establish efforts that promote responsible data use, foster collaboration, and encourage data sharing and reuse, while ensuring that the benefits of genomic technologies for agricultural improvements are shared equitably among all relevant interested parties in the agricultural value chain.

This report explores current practices related to genomic data governance, management, storage, and sharing, and provides insights and actionable recommendations for data standardization. It also identifies challenges and opportunities for data harmonization, and addresses the following key issues:

- Metadata descriptors;
- Data governance and ownership;
- Training for data producers and users; and
- Securing funding for long-term data preservation.

These topics were identified through a literature review and landscape analysis of current best practices for promoting data standards in agricultural genomics (agri-genomics); interviews with internationally recognized researchers responsible for multi-genomic data projects and a genomic data provider; and screening ICT data management plans and analyzing ICT members' responses to a structured questionnaire. An online workshop was also delivered, followed by interviews with relevant parties with expertise in data standardization, research data use, consortium-based industry collaboration, and end-user implementation. These interviews were conducted to better understand perspectives about the relevance, feasibility, and clarity of the recommendations that are proposed in this report. Finally, it should be noted that the scope of this report focuses solely on standardization needs related to methodology and data practices in agri-genomics.

2. Methods

2.1. Literature Review and Landscape Analysis

The literature search strategy identified peer-reviewed articles, literature reviews, and conference proceedings using Google Scholar and online library access. Key search terms related to genomic data, such as "agriculture", "agricultural genomics", "genomic databases", "genomic research", "metadata", "data governance", "data management", and "data practices" were used to identify relevant sources of information, which also included identifying and following pertinent references within articles and whitepapers. The relevance of initial search results and subsequent article evaluations was determined by their relevance to the primary focus of this report, which was to assess opportunities for data standards in agri-genomics. Important topics and trends were synthesized to fill gaps in the current knowledge and identify potential areas for future standards development. The findings were organized into a narrative that highlights both the challenges and areas of existing agreement in genomic data standards, providing a comprehensive overview of this field, existing approaches to implement standards, and future directions impacting agri-genomic data in Canada.

The literature review and landscape analysis identified best practices and processes for data governance and management, as well as data collection, analysis, sharing and reporting. Information sources included not only academic and grey literature but also the practical experience shared by internationally recognized researchers working on three different agri-genomic initiatives. Commonalities and inconsistencies within agri-genomics that affect data interoperability and impede comprehension and comparison between datasets were noted. The literature review and landscape analysis focused on the types of genomic data being generated by the Genome Canada ICTs, which are (a) array genotypes, (b) whole genome sequencing, and (c) RNA sequencing (RNA-seq), at both individual and meta levels.

2.2. Insights from International Interviews

Leaders of three key international agri-genomics initiatives were contacted and interviewed. The primary objective of these interviews was to learn from their experiences and gather insights on genomic data practices and processes that are discussed in this report. The interviewees included a research centre director, a university professor, and a commercial business owner. Each participant was interviewed online, and their feedback and insights were collated and compared. The interviews followed a semi-structured format (see Appendix A), with questions shared in advance and referred to throughout each session. The sequence of questions remained flexible, allowing for additional probing or follow-up questions based on participant responses, and omitting redundant questions when appropriate. A summary of participants' answers and insights on various aspects of agri-genomic data was documented, and all three participants consented to having their responses included in this report.

2.3. Interdisciplinary Challenge Teams

Descriptions of data governance, management, collection, storage and sharing were obtained from six ICTs of Genome Canada's Climate Action Genomics Initiative. This section introduces research projects that are harnessing genomic technologies to address climate challenges in agriculture. Each project targets a unique aspect of the initiative, ranging from soil health and crop resilience to livestock productivity and alternative protein sources. By integrating advanced genomics, data governance, and management alongside interdisciplinary collaboration, the ICTs aim to reduce greenhouse gas (GHG) emissions, improve sustainability, and support the transition to net-zero agriculture in Canada by 2050 [23]. For this report, ICT data management plans were accessed to gather an overview of management strategies, project objectives, targeted genomic data, and technological approaches, along with expected climate benefits, highlighting the ICTs' collective contribution toward Canada's leadership in climate mitigation and agricultural innovation. In addition, a structured questionnaire was shared with ICT members to collect information related to their data practices, as detailed in Section 2.3.1.

2.3.1. Information Collection Matrix

To collect consistent and comparable information across the ICTs, a structured questionnaire was developed (see Appendix B) and shared with all project leaders. The questionnaire served as an information collection matrix, designed to capture key aspects of each ICT's genomic data technologies, data governance and management practices, and climate mitigation strategies and expectations. To provide clear guidance to respondents, the questionnaire was organized into thematic sections that reflect the full lifecycle of genomic data use:

- Project overview and objectives;
- Types of genomic data generated and platforms used;
- Infrastructure availability and outsourcing practices;
- Data governance, management, and sharing protocols;
- Metadata preparation and validation;
- Reporting and communication strategies; and
- End-user engagement and technology adoption.

Each section was designed to capture both technical and strategic dimensions of data governance and management, with an emphasis on data traceability, interoperability, long-term preservation, and accessibility.

Each ICT leader or project manager was asked to provide concise but detailed responses to questions about the types of genomic data collected and generated (e.g., whole genome sequencing, RNA-seq, metagenomics), the platforms used (e.g., Illumina, PacBio, Nanopore), as well as the short- and long-term intended uses of the agri-genomic data. The questionnaire also explored infrastructure availability, including in-house capabilities and third-party services required to support data generation, identifying the challenges related to genomic data pre-processing and quality control, data practices, and computational demands. It is important to note that a significant portion of the questionnaire concentrated on data governance, management, and metadata usage, as the key areas related to the implementation of data standards. Thus, respondents were asked whether any standard operating procedures were already in place, if metadata schemas were implemented as part of ICT routine practices, and how data validation and sharing processes were managed both within and across project teams and institutions collaborating on the same ICT initiative.

The questionnaire also addressed aspects related to long-term data stewardship, including plans for archiving, public accessibility, and alignment with national and international scientific community standards and database repositories requirements, such as Genome Canada's Data Release and Sharing Policies [21] and the National Center for Biotechnology Information (NCBI) databases [24]. ICT project leaders were also asked to identify and report any gaps or roadblocks in their current practices, including lack of funding for long-term data preservation, difficulties in harmonizing metadata within the project or across



national and international collaborators (if applicable), or challenges in integrating multi-omics datasets. All participants that responded to the questionnaire consented to having their input included in this report. The questionnaire's information collection matrix approach provided a rich dataset for comparative analysis, revealing both common practices and areas where harmonization and coordination could enhance the impact and interoperability of genomic research in climate-smart agriculture. In addition, comparing current ICT practices facilitated the identification of existing gaps and opportunities, which informed this report's recommendations for standardization solutions to agri-genomic data in Canada.

2.4. Workshop and Validation

A workshop was held to facilitate discussion and validation of the findings derived from the literature review and landscape analysis, the interviews with international agri-genomic leaders, and the information collected from the ICTs (through both structured questionnaire responses and review of project data management plans). This virtual workshop brought together data scientists, students, researchers, policy-makers, and relevant industry parties. A document providing an overview of the research findings was shared with all participants prior to the workshop to promote discussion during the session. The workshop was structured into three main sections:

1. Overview of CSA Group and the value of standards, followed by introduction to ICT practices;
2. Challenges and recommendations for data tools; and
3. Breakout rooms of small groups for discussion and refinement of proposed recommendations.

Breakout room discussions were guided by predetermined questions, followed by a roundtable group discussion with all participants in the main room. The workshop also provided insights into the different participant perspectives and helped to identify which parties would be most impacted by the adoption and implementation of genomic data standards in Canada. All participants consented to having their responses included in this report.

To further validate the findings from the initial research and the workshops, follow-up interviews were conducted with four participants representing two groups: genomic data users (researchers) or end users of data solutions (industry representatives). These were the two parties workshop participants identified as those who would be most directly impacted by

the implementation of agri-genomic data standards in Canada. The follow-up interview questions (see Appendix D) were developed with the goal of refining standardization needs and addressing the three main hurdles that were identified in the workshop:

1. Metadata templates;
2. Cross-project data practices and governance; and
3. Data ownership/sovereignty and sharing terms.

Using a semi-structured questionnaire, the follow-up interviews corroborated priorities, identified refinements and constraints, and helped refine the recommendations presented in this report. All follow-up interview participants consented to having their responses included in this report.

3. Results and Discussion

3.1. Literature Review and Landscape Analysis

3.1.1. Best Practices and Processes for Genomic Data Collection

As agriculture becomes increasingly data-driven, the collection, documentation, and sharing of genomic information is central to promoting the advancement of genomic technologies and innovations in plant and animal breeding, biodiversity conservation, and sustainable food production [25], [26]. Since 2001, the scope of genomic data has expanded beyond sequencing to include a variety of phenotypes, environmental information, and historical breeding records [27]. Ensuring that the collection of such diverse and complex biological data follows the best practices is essential not only for immediate data use but also for long-term preservation while promoting reusability beyond the original research scope and interoperability with minimum technical and semantic barriers. Moreover, given the complexity of genomic data, recent studies have highlighted a growing consensus that standardized metadata, along with equitable data governance and management are just as critical as the genomic data itself [28], [29].

In a comprehensive systematic review, Chamorro-Padial *et al.* [27] documented the rapid evolution of agricultural data sharing from isolated repositories toward interconnected, artificial intelligence-augmented infrastructures. They also observed an increasing integration of blockchain, internet of things, and cloud-based storage systems to support large-scale agricultural data exchange. However, their findings emphasized persistent structural barriers, especially for smallholder farmers and institutions in developing countries, including inadequate or insufficient digital infrastructure, uncertainty around data ownership, and mistrust in how data might be used by researchers, the broader industry, and political groups. Chamorro-Padial *et al.* argued that targeted investment in infrastructure and capacity building is needed to democratize the benefits of genomic data [27].

Effective data collection depends on FAIR data stewardship [18], which ensures that all types of data are findable, accessible, interoperable, and reusable. Clarke *et al.* [28] concluded that metadata inconsistency is one of the most significant obstacles to implementing the FAIR principles. According to their research, a considerable number of genomic databases either adopt inconsistent identifiers and descriptors or fail to use ontologies, which reduces how easily the data can be found and reused for novel scientific purposes. To address this issue, Clarke *et al.* recommended extensive training on metadata consistency and encouraged research institutions to adopt the use of controlled vocabularies, such as the Crop Ontology, the Animal Trait Ontology, and the Minimum Information about any (x) Sequence (MIxS) [28].

The Genomes Online Database (GOLD) provides comprehensive access to information on genomic and metagenomic sequencing projects [30]. This repository, based in the United States (US), also provides the contextual information associated with each dataset, serving as an international example of metadata consistency. Similarly, the National Microbiome Data Collaborative focuses on advancing microbiome science through data sharing and integration [31],

which provides different types of metadata that can be used to contextualize any experimental data in the microbiome science space, such as sample metadata, preparation metadata, data processing metadata, and feature metadata [32]. These metadata examples can be used to ensure the FAIR principles are effectively implemented from the data collection stage. To support standards and data sharing across existing platforms, both of these initiatives follow community-driven standards, including the taxonomic standards for organisms developed by the Genomic Standards Consortium. These initiatives also aim to leverage data interoperability across different research projects worldwide.

The GOLD repository uses a comprehensive and hierarchical metadata structure that classifies research projects and genomes by their biological source, sequencing method, environmental context, and taxonomic classification [30]. The metadata schema includes more than 180 controlled vocabularies and 5,600 harmonized terms for host organism, geographic origin, sampling protocol, genome completeness, and project purpose (e.g., isolate genome vs. metagenome). These structured metadata entries play a pivotal role in connecting raw data to the plant, microbe, or animal genotype, facilitating downstream interpretations, such as trait mapping, phylogenetic placement, and environmental adaptation analyses. By ensuring that each genomic file is annotated with consistent and detailed context, GOLD enhances data sharing practices and interoperability with other repositories and analytical tools, providing a model for other agri-genomic databases to emulate [30]. Lotterhos *et al.* [29] emphasized the importance of rich spatiotemporal metadata and provenance documentation related to data collection sites in supporting reproducibility and accurate data reuse. Their work highlights tools such as the Genomics Observatories Metadatabase and the ODMAP (Overview, Data, Model, Assessment and Prediction) protocol [33], [34], which allow users to manage and align complex metadata sets that might include GPS coordinates, phenotyping schedules, and environmental conditions, with sample-level genomic data. This tight linkage between sample metadata and genomic output is particularly crucial in plant and livestock systems, where genotype-by-environment interactions are a key component of performance.

Beyond metadata, maintaining high physical sample quality is fundamental to best practices in genomic data collection. The Earth BioGenome Project [35] provides a robust framework for sample collection, recommending harmonized protocols for tissue preservation (e.g., cryogenic freezing) and detailed sample origin records, including voucher identifications [36]. These procedures ensure that the sequence data accurately represents the intended organism and that future researchers can trace its origin. In addition, laboratory-level quality assurance, which includes molecular quantification (e.g., using fluorometers or spectrophotometers), DNA integrity assessments using gel or capillary electrophoresis, and contamination controls, have been widely recommended for high-throughput sequencing workflows, particularly in pathogen genomics applications [37]. Furthermore, to ensure integrity, reliability and interoperability from data collection through to sequence data usage, users must consider the key barriers when integrating data, including long-term funding uncertainty, lack of harmonized data between projects, and lack of collaboration between experts across scientific fields [37], [38].

As agricultural data collection scales up, concerns about security, privacy, and data sovereignty have also grown among researchers and industry parties. Kotal *et al.* [39] proposed policy-aware synthetic data models that allow researchers to share anonymized, yet biologically plausible, genomic data in compliance with the European Union (EU) General Data Protection Regulation and related frameworks [40]. Although this law determines how personal data is processed and protected, it can serve as an example for sensitive agri-genomic data protection incorporated into data governance and management practices. This approach was corroborated by Taddese *et al.* [41], who have called for advanced genomic data governance and stewardship in genomics; FAIR-aligned data management practices; responsible, ethical, and secure handling of sensitive genomic data; controlled access and accountability mechanisms; and balancing

protection of sensitive data with enabling reuse and downstream scientific discovery. These approaches are particularly important in agri-genomics, where genomic evaluations of crops and livestock animals may intersect with commercial interests, national breeding programs, and indigenous genetic resources [42].

Despite the availability of metadata templates following scientific community standards and the FAIR principles, interoperability remains a significant challenge that must be addressed by establishing homogenous operational procedures starting from data collection. Agri-genomics involves not only genetic sequences but also vast datasets on phenotypes, environmental exposures, and management practices. Integrating data from different systems, databases, and platforms along with performance and contextual information requires consistent harmonization across file formats (e.g., FASTA, VCF, CSV, JSON), taxonomies, and units of measurement. Genome Standards Consortium's MIxS is a well-established metadata scientific community standard, which has been broadly adopted to address these challenges. However, Keenum *et al.* [43] reported that a major concern related to reuse remains around data variability, poor documentation, and lack of consistent metadata (including sampling procedures, environment, and project design information), which limits cross-study synthesis and interoperability.

Another aspect of growing importance is the integration of high-throughput phenotyping with genomic data. For example, Sheikh *et al.* [44] reviewed how drone-based hyperspectral imaging and sensor-based high-throughput phenotyping platforms can generate detailed phenotypic profiles for large plant populations. When combined with genomic and environmental information, these data allow researchers to model complex traits such as drought tolerance. Similarly, d'Andrimont *et al. [*45] used convolutional neural networks to classify crop growth stages from street-level images, demonstrating scalable solutions for trait monitoring. These advances further underscore the importance of timestamped, georeferenced metadata, without which it would be impossible to link phenotypic data to specific genetic backgrounds or environmental



conditions. An additional factor to properly connect phenotypic data with genomic information is the establishment of standardized and unique individual sample identification. The International Committee for Animal Recording (ICAR) Global Standard for Livestock Data (ICAR Guidelines) sets guidelines to ensure each animal has a unique and internationally recognized identification that is usable throughout its lifetime [46], [47]. This standard and universal identification allows consistent tracking and integration of genetic data across countries and consecutive years of genomic data generation.

Single-cell transcriptomics are increasingly being applied in agricultural systems. Kapoor *et al.* [48] emphasized that while single-cell approaches offer unprecedented insight into tissue-specific gene expression, their metadata requirements are even more complex than those of bulk sequencing. Protocols for tissue dissociation, cell barcoding, and sequencing depth must be meticulously documented to ensure reproducibility and comparability across experiments. This field will require new scientific standards and data repositories specifically tailored to the unique challenges of single-cell agricultural data. Finally, the role of genebanks and germplasm

collections in genomic data ecosystems is becoming increasingly prominent, emphasizing the need for scientific community standards to be used from genomic data collection throughout the data lifecycle. Anglin *et al.* [49] reported the strategies and best practices that must be considered when genotyping germplasm collections, highlighting the importance of rigorous sample checks with the use of standard passport data (comprised by core set of information describing sample identity, origin, and basic details of the germplasm). These practices boost conservation and allow legacy collections to be mined for important alleles that might be needed to maintain genetic diversity and food security.

As shown in Table 1, the best practices for data collection in agri-genomics now cover every step from fieldwork to database entry and maintenance, integrating harmonized metadata use, secure data governance and management practices, and cutting-edge phenotyping technologies. The GOLD repository stands out as a benchmark for how metadata can be used to contextualize and connect genomic information across different studies, and from genomic data to species-specific and environmental information. However, challenges remain, particularly around interoperability, data ownership, privacy, and equitable access. The literature points to a clear future direction: implementation of metadata standards, including data policies, and infrastructure investments that support both technological excellence and global participation.

Table 1: Summary of required elements for effective data collection.

Data collection effectiveness
FAIR principles
Training on metadata consistency
Adoption of controlled vocabularies
Interoperability across repositories
Structured metadata entries
Protection of agri-genomic data
Integration of phenotyping with genomic data

3.1.2. Best Practices and Processes for Genomic Data Management, Reporting, and Sharing

Advancements in genetic analyses and genomic technologies have significantly increased both the volume and complexity of biological data generated in agricultural research [50], [51]. To effectively maximize the utility and applications of genomic information, it is crucial to establish robust frameworks for data governance and data management, reporting, and sharing [25]. Insights from leading international consortia, social science studies, ethical analyses, and emerging privacy-tech innovations collectively inform a comprehensive approach.

The AgBioData consortium, which comprises experts in agricultural biological databases, data archives, and knowledge bases, was established in 2015 with the goal of identifying common issues in database development, curation, and management. It is also focused on creating database products that are rooted in the FAIR principles [25], [28]. In 2018, the AgBioData consortium published their core recommendations and processes for sustainable genomics and genetics databases that are used in agriculture development [25], which aimed to transform fragmented datasets into an interoperable ecosystem that supports breeding, phenotyping, genetic, and genomic discovery [25]. The recommendations emphasized the importance of standardization, interoperability, persistent identifiers, rich metadata, the use of open and sustainable software platforms, application programming interface (API) accessibility, transparent licensing, regular and clear communication among parties, strategic planning for data sustainability (including diverse and long-term preservation and funding), annotated data with community-controlled vocabularies (e.g., Gene Ontology, Plant Ontology, Trait Ontology) to enable integration and interoperability, and active data biocuration.

Using existing genomic datasets offers several advantages, including saving time and cost, enabling the combination of datasets for meta-analyses and interdisciplinary research, and supporting new discoveries and the exploration of novel hypotheses [52], [53]. From an economic perspective, shared

data has greater intrinsic value, making the more accurate and complete information more valuable for future reuse [26]. Although the importance and benefits of sharing and reusing data have been promoted by international societies, researchers, industry, and funding agencies, the practices and the volume of shared and reused datasets remain low [54]. One reason for this is a lack of common data management practices among researchers working on genomic datasets. This is associated with the preparation of files to be shared, which only include data and metadata relevant to their own project, often without considering whether providing additional data and contextual information might be beneficial to other research efforts or applications.

It is well established that developing a genomic dataset is expensive and time consuming, which often causes researchers to limit the scope of their projects. Thus, adopting best practices for agri-genomic data management and sharing are essential to leverage the use of existing datasets for research purposes and commercial applications. This is possible by implementing, among other practices, "minimal dataset" requirements when preparing metadata and data dictionaries [52]. According to Moore *et al.* [52], this will ensure consistency, utility, and interoperability within and across different datasets, disciplines or locations, making the data useful beyond its original research purposes. A similar approach was previously proposed by Field *et al.* [55], who reported on the importance of providing the minimum information required about genome and metagenomic sequence data, and found that helping researchers and genomic data generators to meet minimum dataset and metadata scientific community standards was considered trivial compared to the cost and time spent generating the data itself. They also found that it reduces redundant efforts and improves data sharing and integration across agri-genomics scientific communities [55].

Moore *et al.* [52] emphasized other key recommendations for advancing best practices in agricultural data management and sharing, including (a) that peer-reviewed publications should be accompanied by adequate metadata detailing collection and analysis methodologies, and (b) that researchers should be encouraged to adopt scientific community standardized collection criteria and use data repositories that ensure long-term preservation. Moreover, funding agencies, scientific journals, and universities should actively promote standards and best practices among researchers, laboratory technicians, and students. In 2023, Clarke *et al.* [28] conducted a survey with all of the databases and resources integrated by the AgBioData consortium to assess the data sharing and ontology needs of the agriculture science community. In total, they received 33 responses from individuals representing 37 genomic, genetic, and breeding databases from both plant and animal data hubs. They found that while API-based access was now widespread, many systems still lacked depth in metadata information, particularly for phenotypic data, and ontology adoption remained inconsistent. They also emphasized that without ongoing training and incentives for data managers and curators, semantic interoperability remains superficial, highlighting the need for tools and education that foster meaningful metadata preparation practices and consistent use throughout a project's duration [28]. Marrano *et al.* [51] supported these findings by introducing a teaching curriculum and training framework designed to advance the FAIR principles. Their curriculum was presented in national and international conferences, aiming to address the gap in educational resources concerning biological databases and their applications for effective FAIR data management within agricultural research.

As a fundamental aspect of best practices in data management and sharing, comprehensive metadata annotation, structured through controlled vocabularies, is essential for facilitating data reuse and interoperability. Genomic databases implementing controlled vocabularies and ontology terms, such as GOLD, Chado, AGROVOC, and Gene Ontology [25], support systems that are both semantically clear and readily searchable. However, the complexity of selecting and applying appropriate ontologies can often discourage users. To address these challenges, the AgBioData consortium has established an education working group and developed an open

access curriculum (AgBioData Curriculum for Ag FAIR Data) encompassing best practices in data management, repository selection, and ontology methodology applications [51]. Reports from this initiative indicate that the educational resources provided to graduate students, professors, librarians, and database curators help lower barriers to adoption and enhance the quality of metadata annotation, which in turn promotes the implementation of FAIR principles and encourages consistent use of scientific community standards among genomic data users [51].

Continuous improvement of metadata guidance and data curation can occur by refining the methods used by researchers and organizations, as well as by maintaining the accuracy and consistency of metadata associated with phenotypic and genomic datasets. This is particularly pertinent for those interested in using large-scale databases (e.g., GOLD, AgBioData consortium) for storage or data reuse as data consumers [30], [56]. For example, the Functional Annotation of Animal Genomes (FAANG) project provides clear and well-documented instructions outlining their metadata principles and structure, which helps ensure that FAANG standards are met when describing experiment design, samples, and analysis prior to incorporating new genomic data into their database [57]. In addition, some essential elements of these large-scale databases include the implementation of standardized metadata formats (e.g., canonical metagenome sample names), the use of controlled vocabularies and ontologies, semi-automated data import processes, and promoting interoperability between databases via data exchange and APIs. Ensuring these industry standards are followed by data providers will support and promote more effective data sharing, integration, and reuse, which contributes to research and collaboration [58].

Although sharing genomic data can facilitate and leverage important health, production, resilience, and climate change discovery benefits, the willingness to share such data is often limited by both practical and psychological barriers. Concerns related to data ownership and privacy from researchers, industry, and producers tend to limit the potential short- and long-term gains for agriculture and society [42]. Gomes *et al. [*59] reported 12 perceived barriers faced by researchers when asked to publicly share their data and codes. These barriers were divided into three groups: (a) knowledge barriers, (b) reuse concerns, and (c) disincentives, which can be respectively characterized by lack of knowledge on how to share data, concern with other users misinterpreting data or using it inappropriately, and fears of being scooped (i.e., when another researcher uses the publicly shared data in a way that the original data collector had planned but not yet completed). Researchers also considered the time investment required to prepare sharable data, and the lack of acknowledgement and recognition for shared outputs. For all perceived barriers, Gomes *et al.* also proposed solutions to promote data sharing, suggesting that integrating metadata into scientific workflows to capture any manual pre-processing steps with sufficient information will help standardize data and ensure it is reusable by others while reducing the workload [59]. Moreover, using dataset DOIs can make data persistent, searchable, findable, and formally citable, and the sharing process provides credit and can be a rewarding part of a researcher's career. This behavioural insight further endorses policy frameworks that explicitly recognize sharing activity in evaluation and funding decisions, as recommended by AgBioData governance discussions [59].

An alternative approach for ethically responsible genomic data sharing was proposed by Byrd *et al.* [58], who advocated a tiered-access model, where public, collaborative, and controlled-access formats are selected based on data sensitivity and risk level. Techniques such as redaction, summary statistics, differential noise, and even synthetic data with clear documentation preserve utility while mitigating disclosure risk (for research and commercial purposes). Importantly, Byrd *et al.* called for structural alignment, which would require that funders, publishers, and institutions establish clear data sharing policies that specify aims and data request procedures (and make data deposition an official requirement), assign DOIs, and integrate sharing behaviours into evaluation criteria, thereby closing the incentive loop and normalizing sharing within scientific culture [58].

Kotal *et al.* [39] and Zafar *et al.* [60] extended these strategies into the agricultural domain by proposing a policy-based, privacy-preserving, synthetic data generation framework. Drawing on EU regulatory requirements (e.g., *General Data Protection Regulation, Artificial Intelligence Act*), Kotal *et al.* generated synthetic datasets that mirror real-world statistical patterns while safeguarding sensitive information [39]. Their tested approach on the ITM4Impact Farm Survey dataset showed minimal loss in predictive power and improved defence against high-risk information and linkage attacks. Crucially, their method enforces privacy rules during data generation rather than post hoc modification, ensuring compliance by design [39]. In addition. Waithira *et al. [*61] reinforces that academic and funding agency policies are critical in shaping sharing cultures. Although this study contains less data than others, its thesis is consistent with current perspectives, which is that establishing coherent policies through metadata harmonization, deposition requirements, timely release, and licensing, can create a framework for structured and legitimate data sharing among institutions and researchers.

A case study in agricultural data management presented an example framework for operational implementation that integrates field sampling, ontology-based metadata capture, cloud archiving, and the engagement of relevant parties [52]. All these elements are driven by supportive policies and sustainable infrastructure that can be applied to genomic datasets. This case study validated the feasibility of the framework and reinforced its scalability, informing a comprehensive model for genomic data management and sharing that incorporates the following six elements:

1. Policy mandates at funder and journal levels, which set baseline expectations for transparency, metadata, licensing, and data sharing behaviour;
2. Technical infrastructure, including APIs, ontologies, persistent identifiers, and synthetic data tools, which enable both compliance and functionality;
3. Education and training, which reduce entry barriers and empower user communities to implement FAIR principles;
4. Cultural incentives, such as DOIs and data citations, which embed recognition for data stewardship;
5. Ethical governance, inspired by Byrd *et al.* [58], which ensures context-sensitive control, privacy awareness, and participant trust; and
6. Resource sustainability through funding, curator support, and iterative improvement, which ensures long-term effectiveness.

When integrated, these elements create a dynamic and iterative ecosystem. Policy frameworks provide essential guardrails, infrastructure supports compliance, education and incentives drive positive behaviour, governance upholds ethical standards and conduct, and sustainability ensures the long-term integrity of the system.

The literature revealed a reinforcing cycle, but challenges remain. The integration of ontology is inconsistent, and the promotion of educational materials relies on principal investigators to motivate team members (e.g., undergrad and grad students, postdoctoral fellows, and laboratory technicians) to develop their knowledge and skills. The translation of policy mandates into tangible researcher practice depends on community investment and infrastructure accessibility. The implementation of privacy-preserving methods, such as those in Kotal *et al.* [39] and Zafar *et al.* [60], requires tool development and integration into existing platforms. Furthermore, the alignment of academic reward structures to credit data stewardship will be crucial for lasting cultural change. Table 2 provides an overview of the required best practices for data management, reporting, and sharing in agri-genomics.

The collective contributions from the AgBioData consortium, behavioural studies, ethical frameworks, policy analyses, and privacy technology research outline a cohesive blueprint for genomic data stewardship in agriculture. In this model, policy, infrastructure, training, governance, innovation, and incentives align to support transparent, secure, and sustainable practices. Together, they have the potential to move agri-genomics toward a future where data is collectively governed, responsibly managed, and widely shared, which will support scientific progress and informed decisions.

Table 2: Summary of required elements for efficient management, reporting, and sharing of agri-genomic data.

Data Management, Reporting, and Sharing Needs
Interoperability
Persistent identifiers
Rich metadata
Open sustainable software platforms
API accessibility
Transparent licensing
Data dictionaries

3.1.3. Best Practices and Processes for Data Analysis

The field of agri-genomics has experienced significant progress, driven by high-throughput sequencing technologies, including genotyping-by-sequencing, whole genome sequencing, and single nucleotide polymorphism arrays. These developments have been accompanied by the introduction of novel protocols, updated policies, new data formats and guidelines, as well as evolving terminologies and ontologies, which are produced by researchers and industry parties and made accessible through data providers and repositories [62]. However, ensuring the accuracy, reproducibility, and usability of genomic data depends on implementing rigorous and standardized pre-processing and quality control practices, which are essential for downstream analyses and data integration [63], [64], [65], [66].

Statistical analyses are typically excluded from operational procedures for genomic data, as researchers often use different methodologies and approaches when analyzing collected genomic information. However, as the importance of the FAIR data principles continues to grow [18], and open, shareable science is increasingly promoted by initiatives such as FAIRsharing [67], maintaining consistent data quality in both schema and values is becoming more critical. Furthermore, it is imperative to adopt integration practices that enhance data quality by prioritizing accuracy and consistency [68].

One of the foundational elements in genomic quality control is the curation and management of metadata. Various studies have stressed the importance of complete, consistent, and harmonized metadata in agri-genomics. Missing or inconsistent metadata can compromise the ability to reproduce and interpret genomic datasets, particularly in multi-institutional or longitudinal studies [62], [69]. The Genomic Standards Consortium [70] and the BioProject initiative [71] have introduced frameworks, such as the MIxS scientific community standards, which provide the minimum information required and structured metadata guidelines for genomics projects [70]. Furthermore, according to the ICAR Guidelines, all metadata provided prior to genotyping should include validated sample identification, collection date, tissue type, and pedigree information [47]. These details should be aligned with consistent ontologies to maintain traceability and compatibility during the verification processes. Agri-genomic initiatives such as AgBioData [25] and GOLD [30] have reported varying degrees of compliance with FAIR principles, particularly in metadata preparation, ontological annotation and sharing, which underscores the need for continued training and short- and long-term institutional support in metadata stewardship. Datasets that are accompanied by metadata containing comprehensive provenance and contextual information about pre- and post-processing steps are more likely to be reusable, interoperable, and trusted by researchers and relevant industry parties.

After metadata has been curated, quality control shifts toward the pre-processing of raw genomic data, properly preparing the database for downstream integration and statistical analyses [62], [66]. The best practices of genomic quality control have several key components, including management of single nucleotide polymorphism quality, assessment of call rates, and filtering based on minor allele frequency and Hardy–Weinberg equilibrium [65], [72]. When working with sequence data, the outputs must undergo stringent quality filtering. Base-call trimming for low-quality Phred scores, removal of adapter sequences, precise demultiplexing of barcodes, and strict verification of restriction-site recognition in genotyping-by-sequencing samples

are essential [73]. After filtering, quality control shifts to alignment, aiming to reduce mapping artifacts that can affect variant calling and genotype imputation [74]. Key steps include removing non-unique and secondary alignments, keeping only properly paired reads, and focusing on mapping quality, depth, and strand bias, which is particularly pertinent in complex or polyploid plant genomes. Tools such as the BWA-MEM algorithm and Bowtie2 enhance mapping accuracy with proper settings [75].

In studies involving structured populations or inbred lines, thresholds are adapted to suit the biological context. For example, a common practice is to remove single nucleotide polymorphisms with a call rate below 90%, a minor allele frequency below 0.05, or those that exhibit significant Hardy–Weinberg equilibrium, unless justified by population structure [65], [66], [72]. Strand bias, allele imbalance, and excessive heterozygosity in inbred populations can signal sequencing errors or alignment issues. For these parameters, agricultural researchers often customize quality control pipelines to account for the unique features of crop and livestock genomes [76], [77]. Another essential component of genomic quality control is reproducibility [78]. It is no longer sufficient to simply report the software tools and parameters used for pre-processing and quality control; instead, researchers are encouraged to package computational workflows in a reproducible and transparent format. The BioCompute Object addresses this need by offering a container for all relevant details of a bioinformatics workflow, including input datasets, software versions, parameter settings, and runtime environments [79]. It was formalized and published as a consensus-based standard under the designation IEEE 2791 [80] to facilitate the reproducibility of results, provide audit trials, and support regulatory submissions. It should be noted that to leverage reproducibility, any pre-processing and quality control implemented in the raw data should be properly documented and integrated to allow future reuse of the genomic data. This is of particular importance when genomic data is collected across several years and distinct projects are further combined to leverage new research projects and knowledge discovery. Defining best practices in analysis is challenging due to changing technologies and varying project goals. For example, reversible or repeatable steps are valuable, filters such as minor allele frequency or Hardy–Weinberg equilibrium can be applied later, and retaining rare variants may be useful. Clearly documenting data processing and preserving information for flexible, non-destructive analysis is essential [65], [66].

Overall, best practices for genomic quality control in agri-genomics span the entire pipeline, starting from metadata curation and raw data pre-processing to alignment filtering, variant validation, assembly assessment, and workflow documentation. The scientific literature consistently emphasized the importance of standard and rigorous quality control in producing reliable, reproducible, and reusable datasets and research outcomes. Thus, the following best practices should be followed:

- Metadata should be complete and aligned to science community standards using controlled vocabularies and ontologies;
- Raw reads should be trimmed, filtered, and checked for contamination;
- Alignments should be evaluated for uniqueness and quality;
- Single nucleotide polymorphisms should be filtered based on coverage, frequency, and biological context;
- Assemblies should include completeness and ortholog-based checks; and
- Workflows should be reproducible and properly documented using harmonized procedures, such as the BioCompute Object.

Collectively, these practices help agri-genomics meet the scientific, ethical, and operational demands of modern data-intensive research. As agri-genomics continues to grow and be integrated into breeding and sustainability programs, adherence to robust and harmonized quality control principles will remain critical to the credibility and impact of genomic discoveries.

3.1.4. Best Practices and Processes for Data Governance

The five primary functions of data governance are [81]:

1. Facilitating data access;
2. Ensuring compliance with relevant national and international agreements and scientific community standards;
3. Supporting appropriate data use and managing potential risks (e.g., data sensitivity and ownership);
4. Fostering equity in access to, use, and analysis of data; and
5. Encouraging the use of data for public benefit.

These functions enhance secure data sharing among industry parties and researchers by establishing operating procedures, polices, and rules for data access while protecting private information as necessary. Data governance is also essential to leverage genomic data availability, providing the necessary underlying justification for local and international data repositories and biobanks [30].

While the FAIR principles are central to data governance, they do not guarantee ethical use or equitable access of agri-genomic data. Consequently, the CARE (collective benefit, authority to control, responsibility, and ethics) principles have emerged as a complementary framework [82]. Rooted in the values of Indigenous data sovereignty and community-led research, the CARE principles are increasingly being applied in biodiversity and agriculture-related genomics databases. For example, when working with traditional landraces or native breeds, applying CARE principles ensures that communities retain agency over their genetic resources and are recognized as rightful partners in the research. This echoes broader trends in human genomics, where the rights and expectations of participants are increasingly central to ethical frameworks. O'Doherty *et al.* [81] argued for a shift from privacy-centric to solidarity-oriented models of governance for human genomic data, proposing a move toward participatory frameworks that embed data usage within broader social values, emphasizing transparency, accountability, and public trust. These recommendations, though framed around human health data, have meaningful parallels in agri-genomics. For example, agricultural data often involves multiple contributors (e.g., farmers, breeders, breed associations, crop producers, growers, commercial laboratories, Indigenous communities, universities, industries, and researchers) each with their own expectations and interests. Applying a solidarity-oriented model in agri-genomics would mean recognizing data contributors not just as sources of material, but also as co-creators of knowledge and future potential knowledge discoveries. This involves developing inclusive governance structures that allow for joint decision-making, shared benefits, and clarity about data use over time.

Regarding data storage, Leigh *et al.* [14] highlighted the best practices for genetic and genomic data archiving and reuse within the fields of ecology and evolutionary biology. Similar to data collection, they identified reusability and interoperability of genomic data as the core challenges, and they proposed recommendations that aligned to their specific fields of study, as well as some general recommendations, including [14]:

- Archive genetic data in raw file formats;
- Archive data within FAIR compliant databases;
- Publicly archive key metadata with the genetic and genomic data;
- Request formalized data management support and a higher value of open data from research institutions, journals, and funding bodies; and
- For work involving Indigenous communities, carefully archive data affected by the CARE principles so data sovereignty is maintained.

This conceptual alignment between human, agricultural, and genomics governance of other disciplines is particularly visible in the growing emphasis on metadata richness, data quality, and interoperability. The challenges of integrating genomic data across platforms and projects are substantial. To address these challenges, initiatives such as the Minimum Information About a Plant Phenotyping

Experiment and harmonized ontologies (e.g., Crop Ontology, Animal Trait Ontology) provide structured vocabularies and annotation protocols [83]. These tools ensure that traits, experimental conditions and design, and data provenance are recorded consistently and meaningfully, allowing datasets to be merged and compared across studies and institutions. For example, FAANG [57] and GOLD [30] not only focus their governance on sequencing but also emphasize extensive metadata curation, including information about animal age, tissue type, housing conditions, and physiological status, while the ICAR Guidelines [47] establish consistent animal identification, promoting accurate data integration across countries for genetic evaluations.

Another core governance challenge is determining how data is accessed and shared. Genomic research benefits from open data, but indiscriminate openness can pose risks, particularly when data is linked to sensitive or commercially valuable information [52]. Tiered access models, where data is shared under different levels of control depending on sensitivity, have become common in both human and agri-genomics data governance rules. In agricultural contexts, data from endangered species or commercially valuable genotypes may require additional safeguards. Research by O'Doherty *et al.* supports the idea that ethical governance does not equate to data hoarding, but rather to responsible stewardship [81]. Responsible sharing, guided by clear data use agreements, time-limited embargoes, and consultations with relevant interested parties, is increasingly seen as best practice.

Legal frameworks also shape data governance in agriculture. The Nagoya Protocol, under the Convention on Biological Diversity, sets international rules for access and benefit-sharing related to genetic resources [84]. Agri-genomics projects that use genetic material from wild relatives or Indigenous communities are expected to comply with the Nagoya Protocol, including obtaining prior informed consent and establishing mutually agreed terms for benefit-sharing. Compliance is not just a legal formality but part of a larger ethical stance that recognizes the historical inequities in genetic resource use. Accordingly, data governance frameworks are shifting to include tools, such as community-developed data licences, data trusts, and collaborative governance boards, that ensure data contributors are included in the process.

Building institutional capacity is another area of focus. In addition to policy, governance also depends on the infrastructure that makes data stewardship possible. Trusted data repositories such as the European Nucleotide Archive and NCBI's GenBank provide secure, long-term storage and assign persistent identifiers to data. Agri-genomics have also developed domain-specific repositories such as Ensembl [85], Ensembl Plants [86], the Animal Quantitative Trait Loci Database (Animal QTLdb) [87], [88], and the Global Agricultural Research Data Innovation and Acceleration Network [89], which integrate genomic data with trait information, environmental data, and breeding histories. These platforms enforce scientific community standards and provide tools for data curation, visualization, and analysis.

Despite the presence of comprehensive international repositories implementing policies and rules for data submission and sharing, data governance issues persist, especially in low-resource environments. To promote equitable participation in global genomics, governance must also address disparities in infrastructure and expertise. This involves investing in capacity building, training local researchers in bioinformatics, and supporting the development of regional data centres. Projects such as the African BioGenome Project explicitly integrate data governance and capacity development to ensure that genomic data generated in Africa remains accessible and beneficial to African scientists and communities [90], [91]. These approaches mirror the governance models proposed in human genomics, which emphasize localized stewardship, benefit-sharing, and capacity investment.

Looking ahead, emerging technologies and the volume of data reflecting the diversity of genomic types will test existing governance frameworks. The integration of genomics, epigenomics (changes in gene expression without any changes to the DNA sequence), phenomics (study of phenotypes), and environmental

data, which can sometimes occur in real time due to advancements in agrotechnology, will require more dynamic governance systems. The rise of machine learning and artificial intelligence technology in agriculture will increase the demand for clean, well-annotated, and interoperable data. Governance must evolve to ensure that such models are transparent, auditable, and free from bias, particularly when they inform breeding decisions or policy recommendations.

The governance of genomic data in agri-genomics is undergoing a profound transformation. As data grows in scale and importance, governance frameworks must balance openness with responsibility, innovation with equity, and efficiency with ethics. The lessons from human genomic governance, particularly the shift toward participatory, solidarity-based models, as proposed by O'Doherty *et al.* [81] offer valuable insights for agriculture. Best practices emphasize alignment with FAIR and CARE principles, integration of metadata standards and ontologies, compliance with legal frameworks such as the Nagoya Protocol [81], [84], use of tiered access models, and investment in infrastructure and capacity building. The objective is to establish a governance framework in which agri-genomic data supports both scientific advancement and societal benefit, promoting innovation that is productive, sustainable, and resilient.

3.2. International Projects

3.2.1. International Interviewees

Semi-structured interviews with key representatives from the following three international agri-genomics initiatives were conducted to gather their perspectives and insights on genomic data practices:

1. **The VistaMilk Research Ireland Centre** [92] helps farmers, food producers, and consumers make better, data-informed decisions for the future of farming and food. The interviewee was a senior principal investigator in quantitative genetics, who is also responsible for the development and implementation of genomic evaluations in sheep, beef and dairy cattle in Ireland. As an academic and strategic leader, the VistaMilk interviewee also has extensive experience in genomic data generation, governance and management.
2. **The Rowan Lab** [93], based at the University of Tennessee, performs cutting-edge research on beef cattle genetics and genomics. The research program is interested in combining multi-omics data and precision technology to identify and breed more efficient, sustainable, and resilient beef cows. The interviewee was a principal investigator of the Rowan Lab, with extensive experience in genomic data collection, governance and management.
3. **The Helical Company Limited** [94] is a private software, database, and services provider company based in New Zealand. The company offers cutting-edge tools for managing and analyzing genetic and genomic data for animal and plant improvement. The interviewee was the owner of Helical.

These international examples were selected because they represent alternative yet aligned (with Canadian examples) initiatives that involve generation, governance, management, analysis, use and deployment of agri-genomic data.

3.2.2. Interviews and Feedback

Sections 3.2.2.1 through 3.2.2.3 provide a summary of the international participants' feedback on different elements of agri-genomic data.

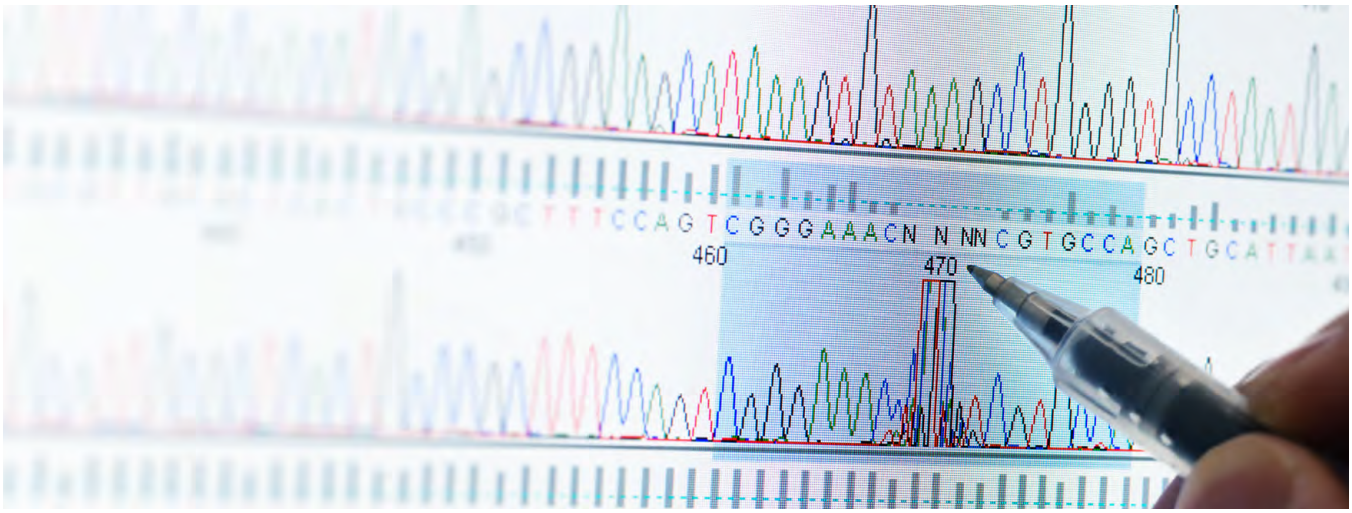
3.2.2.1. Data Collection and Processes

- Genomic data is generated largely by third-party service laboratories but there is also internal capacity (including library preparation), particularly for less routine genomic data types. Array data is used more by industry in agriculture while sequence data, in its many forms, is used more by research in animal and plant disciplines.
- Consideration at the research level is primarily on data collection, and consideration of future use is absent.
- Most genomic data (e.g., array data, sequence data) is usually exchanged in an agreed upon standard format, but the metadata is not.
- Within genomic data types, there are numerous permutations of how the data can exist, so specifying the desired format at the start of a data request or data deposit instruction (e.g., FASTQ vs. BAM files) is encouraged.

- General experience with environmental DNA or eDNA and gene expression data highlighted that data collection tends to be tailored to a specific project, so analysis results may be more shareable and interoperable than the data.
- Raw data pre-processing and quality control are typically conducted using a pipeline that applies standard operational procedures to each data type. Quality control documentation and metadata preparation continue to require additional focus and training for project members. At the service level, a conserved pipeline is used with minimal intervention, generally handled by the bioinformatics specialist.
- Transfer methods of generated data to the end user is inconsistent (e.g., FTP sites, ShareFile, AWS, laboratory information management systems, and physical hard drives). It was noted that generating sequence data often involves issues with DNA extraction, where sequencing difficulties are blamed on DNA quality or quantity extracted. Accuracy of sequence data and storage costs of sequence data over the long-term may also be issues when it comes to more widespread use at the industry level. Transfer of sequence data can be burdensome, with physical movement of hard drives to transfer data still being used.
- Delays in service delivery from third-party laboratories were also highlighted as an issue in genomic data collation, which slows the generation of results from researchers and industry.
- Along with true genomic data, the associated metadata was highlighted as equally if not more important in terms of the usability, interoperability, and longevity of collected data. Nevertheless, metadata preparation remains an area requiring further improvement. Insufficient metadata descriptors and content can decrease the value of genomic data collected, particularly concerning their future reuse.
- Differences in requirements for desired metadata content among groups, including sample naming conventions, is a significant hurdle. Accurate and consistent sample identifiers were frequently mentioned as a key barrier to data exchange and long-term usability. It was described that when samples flow through different entities, it can cause friction and require homework to rejoin samples and genomic data with correct metadata.
- There is a need for alignment between entities to reduce the risk of sample swaps and ensure correct metadata association.
- In some instances, imputation of missing values can be performed to help merge datasets from different sources or time-points.

3.2.2.2. Data Management and Analysis Practices

- Across international initiatives, data management leaders and infrastructure are in place but rely on the commitment of all genomic data contributors to participate, which can lead to inconsistencies and missing information. While uploads to centralized databases are automated as much as possible, novel data emerging from research often poses issues with automation.
- Reluctance to follow data sharing and submission guidelines was a commonly identified theme.
- Unconnected data repositories (e.g., local machines and hard drives) were also described as a problem for ongoing data management and longevity. As part of data management, projects implement different levels of data access and permissions, aiming to protect not only data privacy but also the project's genomic dataset. Although this is an important practice, limited access can lead to delays and incomplete information for end users, impacting the analyses and delivery of results.
- Storage (and analysis) on high performance computing systems were available to all, including cost-effective archival storage, which is particularly used for storing raw data.
- Standard operating procedures for data management were largely absent in the research cases but present when there is an industry service element. However, all three interviewees acknowledged the importance of these procedures, and have personnel adept in data management and analysis, who can build flexible pipelines for multiple data types and sources as needs arise.
- Snakemake and Nextflow were identified as examples of workflow systems used for newly generated genomic data, but new tools are continuously being developed.



- Data analysis leading to quality control of genomic raw files, as well as downstream inference, varied widely in terms of the tools used, pipelines developed, and to a lesser extent, statistical methods used. Although interviewees indicated the use of standard quality control procedures (e.g., available computing infrastructure, skillsets within groups, previous work, and collaborator influence), it should be noted that quality control and downstream analyses are project and data specific.
- Approaches, methods and tools change over time, so standardization may be difficult, even if it was required.
- Only projects or endeavours that have client-facing components have onboarding documents and clear procedures for data management, storage, and governance, ensuring the security and proper use of client or third-party data.
- All participants were familiar with FAIR data principles, and while they may be indirectly achieved, they are not strictly set as a goal and implemented. All participants indicated they were encouraged by this line of questioning to review how they individually manage data procedures.
- Upcoming skills in data management and analysis are drifting from optimum due to the lack of training and reliance on black-box methods. There was emphasis on the need for proper training in data analysis and sharing.
- In both academic and commercial settings, there is onboarding and training (credits) in place but there is still a large element of "learning on the job."

3.2.2.3. Data Reporting, Sharing, and Governance

- All three interviewees indicated that the use of material transfer agreements that outline data use and sharing terms are commonplace. Strict usability terms underpin service agreements with third parties. Data contributors and customers have full control over their data and can share it with third parties. Written permission for data sharing is critical.
- Despite participants having data portals for reporting and centralizing data, researchers frequently hesitate to upload their data due to concerns about data quality, ownership, and the effort involved. This reluctance highlights the need for robust training programs in data management and sharing, standardized metadata practices, and a commitment to maintaining high-quality data inputs, which are currently ad hoc. Ensuring that data management committees are empowered and consistent in their practices can help mitigate these issues.
- Public data repositories (e.g., NCBI) are required to be used by some federally funded research in the US. However, metadata minimum required information and quality is inconsistent on these databases. There are multiple repositories for slightly different genomic data types, such as the Gene Expression Omnibus.
- Isolated data and unoptimized recording and housing of data persists as a barrier to data sharing. Paraphrasing a recurring element from the interviews, the notion of "Excel sheets sitting on hard drives" was described as a common problem for effective data sharing.

- General discussions in the interviews reinforced that communication methods, such as conferences, publications, and technical reports, are essential for reaching diverse parties, including farmers, scientists, and industry members, on the topic of agri-genomic data.
- In terms of data governance, interviewees identified a collaborative approach to data management and sharing as the goal. Governance structures include data management committees that rely on researchers to input data, procedures to ensure data quality, and clear agreement terms for data sharing. In other contexts, end-user agreements contain terms and conditions pertinent to data governance.

3.3. Genomic Technologies and Data from the Interdisciplinary Challenge Teams

Within the six ICTs of Genome Canada's Climate Action Genomics Initiative selected for this research, a range of genomic technologies have been used across beef and dairy cattle, crops, plants, insects and soil microbes, and cells for alternative proteins development. Based on ICT data management plans and the questionnaire responses (see Appendix B) provided by principal investigators and project managers, this section provides an overview of technologies and data practices used by the ICTs, identifying similarities and differences that might impact the implementation of standards for agri-genomic data in Canada.

3.3.1. Interdisciplinary Challenge Teams

This section provides an overview of the six ICTs evaluated for this report.

1. **Climate Action Through Grazing (CAT-G)**

   Canada's grasslands play a critical role in carbon sequestration, supporting biodiversity, sustaining livestock production, and enhancing climate resilience. While these grasslands face increasing pressures, they also have the potential to contribute to Canada's climate solutions and net-zero emissions target [23], and position Canada as a global leader in sustainable grazing and livestock production.

   The CAT-G project is a collaborative initiative that integrates cutting-edge research with on-the-ground grazing management to develop climate-smart grazing strategies that enhance soil carbon sequestration and reduce GHG emissions, while improving beef cattle productivity. The project aims to integrate innovative tools and genomic technologies with traditional practices to address climate change, and its long-term impacts will support both sustainable livelihoods and environmental goals. By protecting and sustainably managing Canada's remaining grasslands, CAT-G's research outcomes include storing an estimated 32 million tonnes of carbon, while delivering multiple benefits, such as enhanced ecosystem health, sustainable beef cattle production, and contributing to both national and global climate goals.

2. **Grassland Genomics for Greenhouse Gas Mitigation (GG4GHG)**

   With more than 16 million hectares of native grasslands and managed forage, the Canadian Prairies offer significant potential for GHG absorption and soil carbon sequestration. The GG4GHG project uses genomics to understand how genetic diversity in these ecosystems impacts their ability to absorb carbon dioxide from the atmosphere and store it in the soil. Specifically, the project will focus on native forage plants, native pollinators, and soil microbes across three Canadian provinces (Manitoba, Saskatchewan, and Alberta) to gain an understanding of the role genetic diversity plays in ecosystem services (e.g., carbon dioxide absorption and carbon sequestration).

   The GG4GHG project works with diverse interested parties, including forage seed producers, cattle farmers, conservation groups, and rural and Indigenous communities, to identify key plant species, understand the ecosystem services they provide, develop predictive analytics, and develop native plant seed mixes that have the greatest potential for mitigating the effects of climate change. The project's main objectives are to:

- Integrate interested party and rightsholder insights into initial project design and survey their constituencies to improve ecosystem service valuation and reduce barriers to adopting beneficial practices, including native species in pasture systems;
- Quantify GHG-related ecosystem service variation across the Prairies and develop genomic indicators to enhance prediction, particularly for carbon sequestration;
- Validate genomic carbon storage indicators across native grasslands and seeded pastures using additional study sites from related team projects; and
- Test project-developed indicators and concepts through field trials and integrate findings into artificial intelligence models and supporting applications.

3. Bio-Inoculants for the Promotion of Nutrient Use Efficiency and Crop Resiliency in Canadian Agriculture (BENEFIT)

Synthetic nitrogen fertilizer production is one of the largest consumers of fossil fuels in agriculture, and its application leads to the formation of nitrous oxide with 300 times the heat trapping potential per molecule compared to carbon dioxide. Canadian agriculture is estimated to account for nearly 75% of all Canadian anthropogenic nitrous oxide emissions, with inorganic nitrogen fertilizer single-handedly accounting for 1.7% of Canada's GHG emissions [95], [96]. Wheat, barley, canola, and dry bean account for approximately 70 to 72% of seeded crop land in Canada [97]. A 10 to 30% reduction in fertilizer application could reduce carbon dioxide equivalent emissions by 0.85 and 2.56 Mt/year, although impacts are likely to be less, assuming a nonlinear relationship between nitrogen application and GHG emissions [96], [98], [99].

The BENEFIT project uses a comprehensive methodology to advance microbial inoculant development for Canadian agriculture. This initiative encompasses microbe isolation, genomic and phenotypic characterization, evaluation of plant responses, and assessment of the economic and environmental impacts associated with the use of inoculants. Its objectives are to implement genomic-driven strategies to identify microbes from Canadian soils with the greatest potential to reduce GHG emissions associated with crop production, and to identify plant functional traits for breeders to generate climate-smart crop varieties that are better able to benefit from microbial inoculants. BENEFIT's outcomes include an anticipated reduction of fertilizer use for cereal (wheat/barley) and brassica (canola/kale) crops of 10% and 30% within five and ten years of project completion, respectively. In addition, the project leaders anticipate a significant reduction, if not complete elimination, of nitrogen fertilization for bean and pea crops within five years of project completion, which is estimated to reduce oil usage by approximately six million barrels per year.

4. Pea Climate-Efficient (PeaCE)

Similar to the BENEFIT project, the PeaCE project aims to reduce the use of nitrogen fertilizer and increase the acreage of nitrogen-fixing pulse crops by developing a climate-resilient, low carbon footprint field pea as a preferred rotation crop using genomic technologies to tackle the most important constraints and challenges facing pea production in Canada. Specifically, PeaCE will harness the genetic diversity of the pea, making it more attractive, profitable, and climate-efficient, as well as more resilient to root rot and drought.

As a multi-institutional and interdisciplinary initiative, the PeaCE project will integrate genomics, breeding, and socio-economic research to develop a stress-resilient, high-yielding pea germplasm with a lower carbon footprint. The project will also generate advanced genomic resources (reference genomes, pan genome assemblies, transcriptomes, proteomes), validate key genes via gene editing, and assess environmental benefits of pea-inclusive rotations. The long-term goals are to accelerate pea breeding pipelines and provide open access resources that benefit producers, industry, and policy-makers globally. Expected outcomes are to increase the adoption of climate-efficient pea varieties, promote decarbonization, and reduce GHG emissions from agricultural soils in Canada, while contributing to economic growth, increased job opportunities, and exportation.

5. Net-Zero Dairy Genome Project (NDGP)

Canada's dairy industry contributes billions to the national GDP and is one of the largest exporters of dairy cattle genetics. To maintain its competitiveness, the dairy industry must adapt its practices to reflect social and environmental values. At the time of writing, dairy production is responsible for approximately 36% of livestock-related GHG emissions in Canada, primarily methane and nitrous oxide. In response, the Canadian dairy industry has committed to a net-zero pledge by 2050, with an interim goal of a 30% reduction by 2030.

The NDGP aims to develop a roadmap for managing GHG emissions in ruminant species by integrating cutting-edge knowledge and datasets from genetics and nutrition [100]. This system-level approach leverages resources created and implemented through previous and ongoing large-scale projects to generate precise emissions estimates at the farm, regional, system, and industry levels. It also identifies efficiencies and opportunities to mitigate dairy cattle enteric GHG emissions. The project is designed to provide accurate, reliable, and robust data that benefit farmers, industry, policy-makers, and GHG inventories related to ruminants. Upon completion, expected outcomes include a 55% reduction in GHG emissions from the Canadian dairy industry, achieved through 30 to 40% improvements in nutrition and 30% through genomic strategies. Project leaders estimate a financial impact of $338 million, along with additional potential reductions in the beef sector. They also anticipate that related genetic advancements in production efficiency will result in an additional $100 million in annual net savings.

6. Omics Guided Technologies for Scalable Production of Cell-Cultivated Meat (Omics4CMeat)

As the global population increases, so does the global demand for dietary protein. Intensive beef farming is a major contributor to livestock-related GHG emissions, and although there are strategies to reduce emissions from livestock production, additional solutions are necessary to satisfy protein needs and achieve carbon neutrality by 2050. The Omics4CMeat project focuses on cell-cultivated meat (also cell-based, lab-grown, or cultured meat) as an emerging alternative protein source. Similar to the production of yogurt and beer, lab-based meat is grown directly from animal cells in a nutrient-rich media inside stainless steel bioreactors. Meat-like textures are achieved by seeding the cells into organic scaffolds, which facilitates the formation of structured products. Prototype lifecycle analyses suggest that gram for gram, cell-cultivated meat generates only a fraction of the environmental impact of conventional beef, including lower GHG emissions, energy consumption, land use, and water requirements. However, to achieve cost parity with animal-based meat, the culturing process must become more efficient, and the culture media must become less expensive.

The Omics4CMeat project aims to use genomics, proteomics, and metabolomics approaches that are underpinned by social science to address technical, economic, and social barriers to the scaling and commercialization of cell-cultivated meat in both Canadian and export markets. Although the project does not specify any anticipated emission reductions, it will address cost and GHG inefficiencies along the production chain to make cell-cultivated meat commercially viable.

These six ICTs focus on distinct biological samples addressed through distinct genomic technologies. Table 3 presents a summary of the biological focus of each ICT project and outlines the range of targets involved. These projects span a range of agricultural systems, from soil and plant health to livestock productivity and alternative protein sources, each contributing uniquely to Canadian climate mitigation targets. Identifying the biological samples helps to inform how genomics is being applied to different components of the agri-food system in Canada, as well as how these applications align with broader goals of sustainability, resilience, and GHG emissions reduction. This overview sets the stage for comparing technological strategies and assessing the collective impact of the ICT initiatives.

Table 3: Summary of biological samples used by each ICT.

ICT	Soil	Plants	Forage	Dairy Cattle	Beef Cattle	Meat Alternatives	Others
CAT-G	X	X	X		X		Grassland ecosystems
BENEFIT	X	X					Microbial inoculants
GG4GHG	X	X	X				Native pollinators and ecosystem services
NDGP				X			
OMICS4CMeat					X (As source cells)	X	
PeaCE	X	X					Microbiome

3.3.2. Potential Benefits of Interdisciplinary Challenge Team Outcomes

The anticipated climate benefits of each ICT outlined in Section 3.3.1 demonstrate a variety of potential contributions to GHG emission mitigation through different biological samples and the application of distinct genomic technologies within the Canadian agriculture sector. Projects such as the NDGP, CAT-G, and GG4GHG, focus on livestock and grassland systems, aiming to reduce methane and nitrous oxide emissions through improved grazing strategies, genomic selection, and ecosystem service enhancement. The NDGP estimates a 55% reduction in dairy-related GHG emissions through genomic and nutritional interventions, while the CAT-G project targets soil carbon sequestration and methane reduction in the Canadian beef industry through microbiome analysis and grazing systems optimization, and the GG4GHG project emphasizes the role of native plant diversity and soil microbes in enhancing carbon storage across the Canadian Prairies.

In terms of crops, the BENEFIT and PeaCE projects focus on lowering emissions by decreasing the use of synthetic fertilizer, which can be achieved by promoting biological nitrogen fixation and developing climate-resilient crop varieties. The BENEFIT project estimates a 10 to 30% reduction in fertilizer application across major Canadian crops, which could lower carbon dioxide emissions by up to 2.56 Mt/year. Similarly, the PeaCE project promotes pea-inclusive rotations that reduce nitrous oxide emissions, with genomic tools being implemented for trait selection and to promote breeding for low carbon varieties. Finally, the Omics4CMeat project offers a novel approach by developing cell-cultivated meat technologies that could significantly reduce the environmental footprint of protein production, but the project's climate impact is more prospective and dependent on future adoption and consumption of lab-grown meat.

While all projects demonstrate high potential for climate mitigation, their value assessments vary based on scalability, readiness, and integration with existing agricultural systems. Projects such as the NDGP and BENEFIT present more immediate and quantifiable benefits due to their alignment with current industry practices and available infrastructure. In contrast, Omics4CMeat faces challenges related to consumer acceptance and commercialization, despite its long-term promise. Common challenges across projects include translating genomic insights into actionable tools, securing funding for implementation, and ensuring uptake by producers and industry. Nonetheless, the collective efforts of these ICT initiatives represent a significant step toward decarbonizing Canadian agriculture through targeted genomic innovation. A summary of each ICT's climate benefits, value assessment, and rationale are presented in Table 4.

Table 4: Summary of anticipated climate benefits for each ICT.

ICT	Climate Benefits	Value Assessment (low, medium, high)	Rationale
CAT-G	▪ Enhanced soil carbon storage ▪ Reduced GHG from beef cattle grazing	High	▪ Direct impact on carbon sequestration and methane reduction in large grassland areas
BENEFIT	▪ Reduced fertilizer use ▪ Lower nitrous oxide emissions ▪ Improved crop resilience	High	▪ Potential for significant GHG reduction across major crop systems
GG4GHG	▪ Increased carbon sequestration ▪ Ecosystem service enhancement	High	▪ Broad ecosystem impact ▪ Supports climate-smart land management
NDGP	▪ Lower methane/nitrous oxide from dairy cattle ▪ Improved breeding	High	▪ Major GHG source targeted ▪ Scalable through industry adoption
Omics4CMeat	▪ Reduced GHG from protein production ▪ Alternative to livestock	Medium-High	▪ High potential, but current market share and adoption are limited
PeaCE	▪ Lower nitrous oxide via biological nitrogen fixation ▪ Resilient crops	High	▪ Direct reduction in fertilizer-derived GHG ▪ Scalable to large crop areas

The anticipated climate benefits also differ among ICTs based on each project's respective degrees of policy and regulatory relevance, with several initiatives explicitly designed to inform national climate strategies, agricultural regulations, and GHG inventories. For example, projects such as NDGP and CAT-G are particularly aligned with regulatory frameworks, as they generate genomic data and predictive tools to support national GHG reporting and mitigation commitments. Specifically, the NDGP is developing genomic evaluations for dairy cattle that include methane emission traits, with the goal of incorporating genomic evaluation into national dairy cattle breeding programs, which will impact both dairy-on-dairy and beef-on-dairy animals while supporting policy decisions. Similarly, the CAT-G project is focused on soil and fecal microbiomes, with the goal of understanding carbon sequestration and methane dynamics in beef cattle grazing systems, which will directly impact land management regulations and climate-smart grazing policies.

Other projects, such as PeaCE and GG4GHG, also have strong and direct policy implications. The PeaCE project supports the Canadian Net-Zero Emissions Accountability Act [23] by promoting pea-inclusive rotations that reduce synthetic fertilizer use and the associated nitrous oxide emissions. The project's genomic data is intended to support variety registration, environmental risk assessments, and carbon credit verification standards, such as the Verified Carbon Standard Sub-program [101] or CarbonClear Canada [102]. The GG4GHG project aims to develop genomic indicators of ecosystem services, such as carbon storage, that can be used by policy-makers to evaluate and incentivize sustainable land use practices in Canada. Although the BENEFIT and Omics4CMeat projects do not express close alignment with specific regulatory frameworks, their efforts will generate valuable data and innovations that have the potential to shape industry standards and inform future policy decisions, particularly regarding microbial inoculant optimization and regulations for alternative proteins.

In general, the six ICTs selected for this report share a common goal to generate genomic data and tools that are both scientifically robust and actionable for regulators and policy-makers. However, the extent of each project's integration with existing regulatory frameworks varies, and only a few reported formal interaction with regulatory bodies during the data reporting phase. This highlights the opportunity to enhance alignment between genomic research outcomes and policy development, facilitating the effective translation of innovations in climate-smart agriculture into regulatory impact.

3.3.3. Potential End Users of Interdisciplinary Challenge Team Outcomes

As shown in Table 5, the ICTs identified a broad range of potential end users for their genomic tools and project outcomes. Targeted end users correspond directly to the associated genomic technologies, which represent an extensive array of applications across soil, plant, livestock, and alternative protein systems. Although there are variations in biological sample targets and research objectives, each initiative was designed to deliver actionable insights and genomic tools to relevant interested parties who are positioned to implement climate mitigation strategies within practical agricultural environments. Common potential end users across ICTs include the different industries, producers, and farmers who are expected to benefit directly from improved and resilient crop varieties, environmentally friendly grazing strategies, or genomic tools for livestock breeding. For example, CAT-G targets cattle producers managing grasslands, PeaCE focuses on pea farmers adopting low carbon rotation crops, while BENEFIT and GG4GHG aim to support growers and livestock producers through microbial inoculants and ecosystem-based genomic indicators.

Table 5: List of potential end users identified by each ICT.

ICT	Potential End Users
CAT-G	▪ Producers (anyone with grazing cattle) ▪ Industry parties ▪ Policy-makers in government or local bodies
BENEFIT	▪ Growers ▪ Companies producing microbial inoculants ▪ Plant breeders
GG4GHG	▪ Livestock producers ▪ Non-governmental environment organizations ▪ Agronomy consultants
NDGP	▪ Industry (breed associations, groups coordinating breeding value calculations) ▪ Basic and applied researchers ▪ General consumers (through improved animal breeding for sustainability, health, productivity)
Omics4CMeat	▪ Industry parties ▪ Policy-makers in government or local bodies ▪ General consumers (public outreach and education about artificial meats technology)
PeaCE	▪ Producers (pea farmers) ▪ Industry parties ▪ Policy-makers in government or local bodies

Industry parties are identified as a key audience for the adoption of genomic tools, particularly for the outcomes from the NDGP and Omics4Cmeat projects. The NDGP's genomic evaluations are designed for integration into dairy cattle breeding programs coordinated by breed associations and industry groups, while Omics4CMeat engages early-stage cellular agriculture companies to scale cell-cultivated meat technologies. Policy-makers are also identified as end users in nearly all projects, with genomic data intended to inform climate policy, regulatory frameworks, and national GHG inventories. Furthermore, some projects, such as PeaCE and CAT-G, identified general consumers as indirect beneficiaries of their genomic tools, especially through public education and sustainability messaging.

While the specific mechanisms for technology transfer vary (from open access databases and applications to publications and field demonstrations), the shared goal among ICTs is to ensure that genomic innovations are accessible, usable, and impactful for those driving change in Canada's agri-food systems.

3.3.4. Interdisciplinary Challenge Team Information Collection Matrix

The ICT screening, which involved the review of data management plans and analysis of questionnaire responses, revealed a range of data practices and challenges among projects. While the ICTs demonstrated a commitment to structured data governance, management, and metadata use, the level of information provided and the practices adopted varied considerably. For example, the NDGP and GG4GHG projects incorporate standards developed by scientific community consortia (e.g., NCBI, Bovine Long Read Consortium Guidelines) for data practices, from generation to sharing genomic information combined with internal and manual metadata review by designated highly qualified personnel. Projects such as CAT-G, PeaCE, and Omics4CMeat rely on centralized systems with external support from a data coordination centre and automated schema for data validation. The BENEFIT project uses metadata templates generated via the schematic engine but lacks designated internal data review protocols. These differences highlight both the diversity of approaches in place and the need for harmonized data practices to support data integration and reuse across Canadian agri-genomic projects.

3.3.5. Interdisciplinary Challenge Team Data Collection, Management and Sharing

Across ICTs, a diverse array of genomic tools and platforms are being used to generate, analyze, and interpret data relevant to climate-smart agriculture. These tools range from high-throughput sequencing technologies to advanced bioinformatics workflows, enabling researchers to explore genetic variation, microbial communities, and functional traits across soil, plants, beef and dairy cattle, and cell-cultivated meat systems. Most ICTs rely on whole genome sequencing, using platforms such as PacBio Revio, Oxford Nanopore, and Illumina NovaSeq, which provide short- and long-read data for genome assembly, as well as identification of genetic variants, including single nucleotide polymorphisms, structural variations, insertions and deletions, and metagenome analysis. For example, the PeaCE project uses a hierarchical sequencing strategy combining multiple platforms to assemble complex pea genomes, while NDGP applies PacBio high-fidelity reads to identify structural variants in dairy cattle. RNA-seq is also widely used, particularly by the BENEFIT, PeaCE, and Omics4CMeat projects, to study gene expression, stress responses, and cell differentiation. Uniquely, Omics4CMeat applies single-cell RNA-seq via 10x Genomics to identify optimal cell types for cell-cultivated meat production.

In addition to whole genome sequencing, CAT-G, GG4GHG, and BENEFIT use amplicon metagenomic sequencing to profile microbial communities in soil and fecal samples, using platforms such as Illumina and PacBio. These approaches are complemented by digital polymerase chain reaction (PCR) in the CAT-G project for methane marker identification and further quantification, while the PeaCE project uses optical mapping and Hi-C scaffolding for chromosome-level genome resolution. Data analysis is supported by cloud-based workflows (e.g., Nextflow, Workflow Description Language), custom pipelines (e.g., Snakemake in NDGP), and quality control tools such as BUSCO, RepeatMasker, and Verkko. Together,

these genomic tools enable high-resolution insights into biological systems and provide the foundation for developing climate-resilient agricultural technologies.

Table 6 presents a summary of how each ICT project approaches data collection, management and sharing. These dimensions are critical for understanding the operational infrastructure behind each ICT and how genomic data is being handled across diverse agricultural contexts in Canada. The table also highlights which platforms are being used and whether there are dedicated data management personnel or protocols in place. Comparing these practices provides insight into the strengths, gaps, and opportunities for harmonizing data standards across projects to support interoperability, reproducibility, and broader impact.

Table 6: Summary of data collection, management and sharing practices among ICTs.

ICT	Data Collection	Data Management	Data Sharing
CAT-G	▪ Amplicon metagenome sequencing, digital PCR ▪ Soil/fecal samples ▪ Illumina, PacBio	▪ Internal database ▪ Standard operating procedures ▪ Group-based access ▪ Data coordination centre support	▪ Internal sharing ▪ Future public release ▪ Project website
BENEFIT	▪ Whole genome sequencing, RNA-seq, metagenome ▪ Soil/plants ▪ Illumina, Nanopore, PacBio	▪ Principal investigator managed ▪ Standard operating procedures ▪ Regular backup on external hard drives and to the Digital Research Alliance of Canada	▪ NCBI ▪ Open access ▪ GitHub ▪ Publications
GG4GHG	▪ Whole genome sequencing, metagenome, array ▪ Plants/soil ▪ Illumina, Nanopore, PacBio	▪ Standard operating procedures ▪ Dedicated reviewers ▪ Institutional storage	▪ NCBI ▪ Internal ▪ Processed data ▪ Publications
NDGP	▪ Whole genome sequencing, array ▪ Cattle ▪ PacBio Revio	▪ Standard operating procedures ▪ Snakemake workflow ▪ Backup to Digital Research Alliance of Canada and on offline storage media	▪ NCBI ▪ Digital Research Alliance of Canada ▪ Industry meetings
Omics4CMeat	▪ Single-cell RNA-seq ▪ Muscle cells ▪ 10x Genomics	▪ Standard operating procedures ▪ Snakemake workflow ▪ Backup to Digital Research Alliance of Canada and on offline storage media	▪ Manuscripts ▪ Internal ▪ Future NCBI
PeaCE	▪ Whole genome sequencing, RNA-seq, Hi-C, optical mapping, metagenome ▪ Peas/soil ▪ PacBio, Nanopore, Illumina	▪ Internal database ▪ Standard operating procedures ▪ Data coordination centre support ▪ Cloud pipelines	▪ NCBI ▪ Project dashboard ▪ Publications

A variety of challenges in generating genomic data were observed, particularly around consistency, accuracy, and quality. One common issue across several projects was variability in sample quality, particularly when working with environmental samples such as soil and feces. CAT-G, GG4GHG, and NDGP indicated that low-quality samples could compromise their sequencing depth and data integrity, sometimes requiring re-sequencing, leading to additional cost and time. The PeaCE project also highlighted the difficulty of working with large and repetitive genomes, which complicates assembly and increases computational demands. Additionally, batch effects across sequencing runs and inconsistencies in library preparation were cited as factors that reduce reproducibility and comparability across datasets. Despite the shared concerns, the projects differed in how they address these challenges. NDGP and GG4GHG have implemented quality control protocols and pilot sequencing runs to optimize data generation, while PeaCE uses a hierarchical sequencing strategy and specialized software (Verkko and BUSCO) to improve genome assembly. CAT-G and GG4GHG emphasized the need for robust bioinformatics infrastructure to handle large metagenomic datasets, while Omics4CMeat faced unique challenges in analyzing single-cell RNA-seq data due to limited expertise and lower RNA quality from fresh cells. Notably, BENEFIT reported no major challenges in data generation, suggesting a more streamlined or controlled workflow. These differences reflect the diversity of highly qualified personnel required, biological samples targeted, and sequencing strategies implemented across projects, and underscore the need for shared standards and collaborative solutions to improve data quality and consistency in agri-genomics. In addition, the ICTs demonstrated a range of practices in data management and sharing, shaped by their organizational structures, technical capacities, and national and international collaboration models. Most projects have established internal systems for managing genomic data, with varying degrees of centralization and oversight. CAT-G, PeaCE, and Omics4CMeat benefit from centralized infrastructure provided by the ICT data coordination centre, which enables researchers to upload, view, and export data across teams with external data specialist support. These systems support regular updates and controlled access, with principal investigators and key personnel overseeing data governance. In contrast, BENEFIT and GG4GHG operate under frameworks led by principal investigators or highly qualified personnel (e.g., postdoctoral fellows, bioinformaticians, and laboratory technicians), where data management is more decentralized and embedded in project-specific workflows. The NDGP stands out for its integration with the Bovine Long Read Consortium [103], which provides guidelines describing data requirements and steps to process sequence reads followed by external review protocols.

In terms of data sharing, all projects expressed a commitment to transparency and open science, following the FAIR principles, with plans to deposit genomic datasets in public repositories such as NCBI's Sequence Read Archive, GenBank, and Gene Expression Omnibus, which are all US federally funded genomic data repositories built and maintained by NCBI. Specifically, Sequence Read Archive is a raw sequence data repository, GenBank is an annotated nucleotide sequence repository, and Gene Expression Omnibus is a functional genomics (gene expression and epigenomics) repository. BENEFIT, GG4GHG, and NDGP have already aligned their practices with these platforms, while CAT-G, PeaCE, and Omics4CMeat are in the earlier stages of external sharing. Internal sharing policies varied, with some projects restricting access to group-specific data, while others are allowing broader access among national and international collaborators. NDGP and GG4GHG have more formalized sharing protocols, including permission-based access and data backup strategies. It is important to note that long-term data preservation remains a challenge across all ICTs, with most lacking dedicated funding or infrastructure to ensure sustainability beyond the project lifecycle. Metadata management is a common strength among the projects, with ICTs reporting the use of structured schemas to ensure traceability and consistency throughout their duration. These schemas typically include sample identifiers, sequencing parameters, and contextual information, such as collection site or experimental conditions.

NDGP and GG4GHG incorporate manual review processes to ensure and validate metadata accuracy, while CAT-G, PeaCE, and Omics4CMeat rely on automated schema comparison tools. BENEFIT uses the schematic engine to generate metadata templates but has not designated a reviewer who is responsible for metadata validation. Despite these efforts, harmonizing metadata across collaborators and platforms remains a challenge, particularly in projects involving multiple sequencing technologies or multiple national and international partners.

Overall, while the projects share a commitment to robust data management and open sharing, they differ in their levels of formalization, technical infrastructure, and long-term planning. The use of public repositories and structured metadata reflects a collective understanding of best practices, but gaps in funding, data practices harmonization, and interoperability highlight the need for coordinated national efforts.

3.3.6. Interdisciplinary Challenge Team Governance and Metadata

ICT data governance and metadata practices reflect a blend of centralized coordination, project-specific protocols, and emerging scientific community standards. Most ICTs rely on internal governance structures led by principal investigators or supported by the data coordination centre, which provides infrastructure for secure data storage, controlled access, and collaborative sharing among institutions and team members. CAT-G, PeaCE, and Omics4CMeat benefit from centralized systems that allow researchers to upload, view, and export data across teams, while BENEFIT and GG4GHG operate strictly under frameworks managed by principal investigators and highly qualified personnel, with varying degrees of formal oversight. NDGP stands out for its integration with the Bovine Long Read Consortium, which enforces guidelines for data workflow and external review of datasets and metadata. In this case, a dedicated team member (highly qualified personnel) is responsible for implementing the Snakemake workflow for data pre-processing and further conducting data quality control analysis, as required by the Bovine Long Read Consortium subcommittee.

Despite these differences, ICTs demonstrate a commitment to establish structured metadata practices, ensuring traceability and consistency as integral components of their data governance strategies. These schemas typically include sample identifiers, sequencing parameters, and contextual information, such as collection site or experimental conditions. The NDGP and GG4GHG projects incorporate a standard operating procedure and a homogeneous metadata template that dictates the file names used for genomic data, which will be in place throughout the project lifecycle. Specifically, the NDGP metadata template requires file names to be the same as the NCBI BioSample identifier or the international identifier, along with animal breed and sex, sequence technology, mapped coverage, read N50, average read quality. If an animal's sequence has not yet been submitted to a data repository or is missing the BioSample or international identifier, an identification will be created following the Interbull animal identification rules [47], which consists of the 3-letter breed code, 3-letter country code, sex code (M or F), and 12-character animal identification. If sire and dam identifiers are available, they should also follow the same rules on animal identification, as proposed by the ICAR Guidelines [47].

According to NDGP's data governance practices, although the NCBI BioSample identifier does not have to be filled in at the time of metadata generation, it is required to be added following submission to the NCBI website and prior to publication. Moreover, when single nucleotide polymorphism chip data is available, this information is also indicated in their metadata. Other ICTs, such as CAT-G, PeaCE, and Omics4CMeat, rely on manual review processes to ensure and validate metadata accuracy. For these projects, the groups create data schemas specific to each dataset, including the data type and relevant metadata fields, which can vary between different data types. When the data is uploaded to the data coordination centre, it is directly compared against the corresponding schema as part of their data validation process. To ensure the metadata files are not corrupted and contain the appropriate information, data validation is performed automatically

Table 7: Summary of ICT governance and metadata strategies.

ICT	Governance Lead	Metadata Review	Long-Term Governance
CAT-G	Data coordination centre	Automated	Goal, not funded
BENEFIT	Principal investigator	None	Upload to NCBI only
GG4GHG	Highly qualified personnel	Manual and standard operating procedures	Not funded
NDGP	Dedicated team	Manual (Bovine Long Read Consortium)	Digital Research Alliance Canada and NCBI
Omics4CMeat	Data coordination centre	Automated	Goal, not funded
PeaCE	Data coordination centre	Automated	Goal, not funded

by comparing uploads against the defined schema. BENEFIT uses the schematic engine to generate metadata templates but has not designated a project member as reviewer for data validation. Although governance practices and metadata schemas are widely used, achieving harmonization among ICTs and within ICT groups remains difficult. The harmonization of data practices is particularly difficult due to the use of various sequencing providers, the biological samples targeted, and the high diversity of project collaborators contributing not only with new data but also, in some cases, with historical data. Table 7 provides a summary of how each ICT project approaches data governance and metadata practices

At the time of writing, the three main challenges associated with data governance reported across ICT projects are:

1. Insufficient funding and infrastructure for long-term data preservation;
2. Inconsistencies and limited comparability of metadata formats among collaborators and platforms; and
3. Limited harmonization within and across teams.

ICTs often face difficulties in aligning file naming conventions, reference genome versions, and variant calling pipelines, particularly when working with multi-omics data or diverse biological targets. While most projects align with Genome Canada's data sharing policies and the required use of public repositories for data storage (e.g., NCBI, Sequence Read Archive, and Gene Expression Omnibus), the absence of Canadian data practices standards and centralized oversight limit the scalability and reusability of agri-genomic data. In addition, the lack of formal certification or external auditing can affect data quality and trustworthiness, nationally and internationally. These challenges highlight the need for coordinated efforts to establish national harmonization and standards for genomic data governance and metadata management, enabling more effective collaboration, data integration, and long-term impact across Canada's climate-smart agriculture initiatives.

3.3.7. Interdisciplinary Challenge Team National and International Uptake

The ICTs funded by Genome Canada show varying levels of national and international uptake of the genomic technologies they are developing, reflecting

differences in maturity, scalability, and alignment with existing agricultural systems, as outlined in Table 8. The NDGP and BENEFIT projects are closely integrated with Canadian industry and research networks, making their technologies more immediately applicable to dairy cattle producers and growers, respectively. Specifically, NDGP's genomic evaluations for methane emissions in dairy cattle are designed to be adopted by breed associations and industry groups (e.g., Lactanet), which already play a central role in national dairy cattle genetic evaluations. Similarly, BENEFIT's microbial inoculant strategies are aligned with crop production systems widely used across Canada, and the project's open science approach facilitates rapid dissemination and uptake by growers, plant breeders, and inoculant producers.

Internationally, several of the ICTs are building on or contributing to global genomic initiatives. For example, PeaCE collaborates with the US Department of Agriculture, the International Center for Agricultural Research in the Dry Areas, and the Australian Grains Genebank, and contributes to the Pisum 60+ International Pea Pan Genome Sequencing Consortium, positioning its genomic resources for global use in the space of pea breeding. Similarly, GG4GHG and CAT-G align with international efforts to quantify and enhance ecosystem services through genomics, with parallels in grassland management programs in New Zealand, Australia, and Europe. These projects are generating genomic indicators and predictive tools that could inform sustainable land use policies and conservation strategies beyond Canada.

Table 8: Summary of national and international uptake for ICT genomic technologies.

ICT	National Uptake	International Uptake	Notes
NDGP	Integrated with Lactanet and national breeding programs	Aligned with genomic selection efforts in EU, US, and Oceania	Strong industry alignment; scalable through existing infrastructure
BENEFIT	Open science model; accessible to growers and breeders	Microbial inoculants widely used globally; genomics adoption growing	Streamlined workflows; metadata risks if not properly managed
PeaCE	Connected to Canadian pea breeding networks	Collaborations with US Department of Agriculture, International Center for Agricultural Research in the Dry Areas, Australian Grains Genebank, and Pisum 60+ Consortium	Strong global relevance; challenges with genome complexity
GG4GHG	Supports prairie land management and conservation	Parallels with EU and US ecosystem genomics programs	Multi-species focus; integration challenges across platforms
CAT-G	Aligned with Living Laboratories and sustainable ranching	Similar approaches in Australia, New Zealand, and Europe	Large data volumes and variability in sample types
Omics4CMeat	Emerging field; academic-industry consortia forming	Rapid growth in US, Israel, and Singapore	High potential, but commercialization and consumer acceptance are hurdles

The Omics4CMeat project represents an emerging area of genomic application, with its focus on the development of cell-cultivated meat technologies. The cellular agriculture industry, within which Omics4CMeat conducts its research project, has grown considerably since 2021, when it attracted over US$9.7 billion in global investments, and there were more than 400 companies active in this sector in 2025 [104]. While uptake in Canada is still in the early stages, Omics4CMeat is part of a growing global movement for the development of alternative protein products, with advancements in countries such as the US, Israel, and Singapore [105], [106]. The integration of single-cell RNA-seq and proteomics enables the project to advance national and international scientific community standards and innovation pipelines for alternative protein production. However, broad adoption in Canada will require regulatory approval and positive consumer acceptance.

Overall, all projects demonstrate strong potential for both national and international uptake, particularly where genomic tools are designed to integrate with existing breeding programs, production systems, or global research networks. Common enablers of uptake include open access data sharing, alignment with policy goals (e.g., Canadian Net-Zero Emissions Accountability Act), and collaboration with industry and international partners. However, challenges remain in translating genomic insights into scalable technologies, securing long-term funding, and harmonizing data community standards to support cross-border interoperability.

Based on the semi-structured interviews with key representatives from three international agri-genomics initiatives (VistaMilk, Rowan Lab, and Helical), there is significant overlap in the types of genomic technologies being used and discussed, many of which are also central to the ICTs. All three interviewees referenced the use of genomic sequencing technologies, such as whole genome sequencing, single nucleotide polymorphism arrays, and gene expression data (RNA-seq). VistaMilk and Rowan Lab indicated that they are generating sequence data internally using platforms like NovaSeq. In addition, these research groups are also working with array-based genotyping platforms for parentage tests and breeding value estimations. Similarly, Helical software systems are designed to handle array data, sequence data, and gene expression data, while providing infrastructure for genomic data storage, processing, and evaluation, mirroring the data types and workflows reported by ICTs such as NDGP, GG4GHG, and PeaCE.

The international interviewees also emphasized the progress and persistent challenges related to integrating and managing diverse genomic data types, as well as the adoption and integration of genomic technologies at both research and industry levels. DNA extraction quality, sequencing accuracy, and the complexity of handling large data volumes are challenges identified by both the interviewees and the ICTs. There is a shared recognition that while array data is widely adopted in commercial breeding and industry settings, sequence and gene expression data are still primarily used in research due to higher costs, data management complexity, and the need for clear economic benefits for end users. This is consistent with the ICT questionnaire responses, where array genotyping was identified as a mainstay for livestock and crop improvement, while whole genome sequencing and RNA-seq are more often used for discovery and trait mapping.

The importance of data practices harmonization and data sharing infrastructure were key recurring themes. The international interviewees confirmed that while there is growing recognition of the need for consistent data governance, management and sharing, similar to the approaches used in the NDGP, GG4GHG, and CAT-G projects, actual implementation remains inconsistent. For example, VistaMilk has established data portals and management committees, but uptake is hampered by researcher reluctance to upload data, concerns about data quality, and the absence of dedicated data management personnel. Similarly, the Rowan Lab interviewee underscored the challenges of aligning metadata and naming conventions across different laboratories and data sources, echoing the metadata harmonization issues reported by PeaCE and GG4GHG.



"To address the need for high-quality data, future efforts must focus on the development and adoption of harmonized and consistent protocols for data generation, quality control and pre-processing, and validation."

Finally, the interviews with international experts also emphasized the critical importance of training, effective data governance, and user-friendly systems to facilitate the adaptation of genomic technologies at both national and international levels, core themes that were also identified by the ICTs. Both academic and industry parties highlighted the need for clear and consistent operating procedures, robust metadata community standards, and ongoing training to ensure data quality and reusability. The lack of dedicated data managers and the reliance on researchers for data entry and curation are seen as barriers to effective data sharing and long-term data legacy. These findings align with the challenges reported by the ICTs, reinforcing the need for coordinated national and international efforts to establish harmonized data practices, support infrastructure, and foster a culture of data stewardship in agri-genomics.

3.3.8. Existing and Future Challenges for Interdisciplinary Challenge Team Data

3.3.8.1. Data Consistency and Quality

Ensuring data consistency and quality is a critical priority for genomic research, particularly in large-scale, multi-institutional projects. As genomic datasets grow in complexity and volume, it becomes increasingly important to maintain consistency across sequencing platforms, biological sample types, and analytical pipelines. Consistent data practices enable integration across studies, leading to reliable comparisons and reproducibility of results, which is essential for translating genomic insights into actionable agricultural innovations. High-quality data also underpins the credibility of findings, particularly when informing breeding decisions, environmental assessments, or policy recommendations.

To address the need for high-quality data, future efforts must focus on the development and adoption of harmonized and consistent protocols for data generation, quality control and pre-processing, and validation. This includes implementing robust quality control measures at every stage, from sample collection and DNA extraction to sequencing and downstream analysis. Agri-genomic projects could adopt common reference genomes, unified variant calling pipelines, and standardized file naming to minimize variability and support data integration beyond the project lifecycle. Additionally, the use of containerized workflows and workflow management systems could help ensure reproducibility and transparency in data processing.

Equally important is the consistency of metadata, which provides the contextual information associated with genomic studies that are necessary to interpret genomic data accurately, as well as to leverage data reusability and improve the ability to reproduce results [43]. According to the Genomic Standards Consortium,

contextual metadata descriptors must include clear environmental information (from study design to sample collection site details) and genomic specific MIxS, establishing common metadata schemas with minimal information about the data, consistent and controlled vocabularies, and validation tools, all of which are essential for ensuring that datasets are interoperable and reusable during and beyond the project lifecycle. Moreover, research funders and project leaders should also invest in training researchers and highly qualified personnel on metadata best practices, educating the upcoming generation of researchers, and consider assigning dedicated data stewards to oversee metadata curation and compliance. By aligning with national and international tools and technical standards of scientific and industry communities, such as those promoted by Genome Canada, GOLD, the Global Alliance for Genomics and Health, and the FAIR principles, ICTs can enhance the long-term value and impact of their genomic data to promote GHG emission reductions across the Canadian agriculture sector.

Ultimately, enforcing metadata use while enhancing data consistency and quality extends beyond technical considerations, encompassing organizational and cultural dimensions as well. Achieving these improvements requires collaborative and coordinated efforts among institutions, funding agencies, and researchers. Prioritizing data stewardship, incentivizing the adoption and use of best data practices, and integrating quality assurance throughout the research project lifecycle and beyond are essential steps. As genomic technologies continue to evolve and expand in agricultural applications, such coordinated efforts are crucial for maximizing the scientific, economic, and environmental benefits of genomic data.

3.3.8.2. Metadata and Standards Alignment

Standardized genomic and metadata practices enable ICTs to translate research outputs into usable formats for policy-makers, regulators, industry, and the public. Across projects, data harmonization is the key enabler of interoperability and reuse, particularly in metadata, persistent identifiers, and the documentation of data quality control and pre-processing. Adopting shared scientific community standards, such as MIxS and the Crop Ontology, ensures that genomic and phenotypic datasets are consistently described. The routine use of persistent identifiers, including DOIs for datasets and Open Researcher and Contributor ID (ORCID) for contributors, links data products with the people and publications generated throughout a research project. Together, these practices enable long-term traceability, facilitate integration across ICTs and with external collaborators, and reduce redundancy in data collection.

3.3.8.3. Data Lifecycle Management and Training

To maintain high-quality and reusable data, each ICT should formalize and adopt consistent operating procedures that describe data handling from acquisition through archiving. Dedicated data curators should monitor compliance with metadata and submission requirements. Training programs are necessary to strengthen data curation, quality control, and schema-based uploads. Automated validation routines and controlled vocabulary should be implemented to minimize manual inconsistencies. These measures support long-term continuity, allowing datasets to remain usable and well-documented even after individual projects conclude. This combination of data practices can then be used to drive data standards for agri-genomic projects in Canada.

3.3.8.4. Data Presentation for Decision-Making

To ensure genomic data effectively drives decision-making, research genomic data and outputs must be tailored to meet the specific requirements of different parties, including:

1. Regulators and policy-makers, who require high-level indicators that summarize genomic diversity, environmental impact, and data coverage. Interactive dashboards and concise policy briefs can translate complex datasets into actionable insights for sustainable agriculture policies.
2. Industry partners, who benefit from benchmark reports and trend analyses that relate genomic findings to productivity or resilience outcomes. Integrative repositories that combine genomic, phenotypic, and environmental data can guide breeding and investment priorities.

3. Researchers and scientific users, who benefit from well-documented, consistent datasets that can be integrated across projects and reused for meta-analyses or model validation. Clear data provenance associated with harmonized metadata and file formats allows researchers to efficiently build on previous research project outputs, reducing duplication and promoting reproducibility.
4. Consumers and the general public, who require accessible narratives and visualizations that connect genomic research to food quality, traceability, and climate resilience. Public-facing summaries strengthen understanding and trust in agri-genomics.

3.3.8.5. Policy and Regulatory Impact of Interdisciplinary Challenge Team Data

An important emerging challenge across the climate-smart agriculture ICTs is the limited awareness among researchers regarding the potential policy and regulatory impact of their genomic data and project outcomes. While all of the projects discussed in this report generate high-quality, policy-relevant data, such as genomic indicators for GHG emissions, soil health, or crop resilience, few have formal mechanisms in place to translate these findings into actionable insights for policy-makers. This suggests a need for more intentional engagement with regulatory parties and a clearer understanding among researchers of how their data can inform national and international climate and agricultural policies.

For example, Omics4CMeat, which focuses on developing scalable cell-cultivated meat technologies using single-cell RNA-seq and proteomics, has significant implications for both policy-makers and consumers. As governments begin to consider regulatory frameworks for alternative proteins, the data generated by this project could inform food safety standards, labelling requirements, consumer acceptance, and sustainability assessments. However, the lack of a direct pathway for engaging with regulatory bodies may delay the translation of project findings into policy. Similarly, PeaCE and BENEFIT generate genomic data that could support fertilizer reduction targets and carbon credit systems, yet their focus remains on scientific outputs rather than regulatory integration.

The NDGP and CAT-G are more explicitly aligned with policy objectives, such as improving GHG inventories and supporting the Canadian Net-Zero Emissions Accountability Act. The NDGP is developing genomic evaluations for methane emissions in dairy cattle, which could be integrated into national breeding programs and used by regulatory agencies before project completion. However, even in these cases, the pathway from data generation to policy adoption is not always clearly defined, and researchers may not fully recognize the importance of tailoring their outputs to meet regulatory requirements or decision-making needs.

To enhance the policy and regulatory impact of genomic research, there is a need for greater awareness and capacity building among researchers and funding agencies. This includes training on how to communicate findings to policy audiences, developing policy briefs or technical guidelines beyond manuscript publications, and engaging with regulatory parties early in the research process. Establishing formal mechanisms for knowledge mobilization and policy translation is essential to ensure that the valuable genomic data produced by agri-genomic projects in Canada contributes meaningfully to shaping sustainable agricultural policies and climate action strategies.

3.4. Use Cases for Standardized Interdisciplinary Challenge Team Data

There are three main use cases that illustrate the practical benefits of standardization for Canadian agri-genomic innovation:

1. **Policy and Regulatory Evaluation**

   The adoption of uniform metadata and controlled vocabularies enhances regulatory transparency by effectively connecting genomic datasets to their source of origin, study designs, sampling methods, and analytical context. When datasets are annotated with persistent identifiers (e.g., DOIs or

ORCIDs) and uniform descriptors (e.g., MIxS and the Crop Ontology), regulatory agencies can rapidly and effectively track relevant data information, see how genomic data was submitted, and assess the anticipated environmental impact of adopting new genomic tools. This capacity is essential for cross-jurisdictional policy evaluation, ensuring data comparability among provinces, as well as between Canada and international frameworks, including the Global Alliance for Genomics and Health.

2. **Breeding and Industry Application**
   Uniform identifiers and interoperable repositories facilitate the integration of datasets obtained across multiple years, locations, and species for breeding programs and industry parties. For example, the use of uniform and consistent sample naming conventions along with validated metadata support genomic-phenomic association, which is essential for advancing genomic selection pipelines in crops, livestock, poultry, and aquaculture. This approach reduces redundant sequencing efforts and streamlines decision-making processes for cultivar or breed release, aligning with the FAIR principles of data reuse.

3. **Public and Scientific Transparency**
   Applying metadata templates, persistent identifiers, and repository integration ensures that project outputs remain findable and accessible beyond a project's lifetime. These practices, which are common in the US Department of Agriculture's Ag Data Commons, illustrate how a government-run platform can maintain long-term discoverability of datasets across disciplines. Appendix C discusses the Ag Data Commons as a model of public agricultural data management and its relevance to ICT long-term data stewardship. Through such models, data records are curated, assigned DOIs, and linked to associated publications, allowing users to move between research outputs and the underlying datasets. Adopting similar practices across the ICTs would ensure that secondary users, including researchers, policy analysts, and industry partners, can locate and reuse data long after individual projects end. This bidirectional connection between datasets and peer-reviewed papers enhances accountability and strengthens public trust in genomics-based agricultural innovation.

3.5. Roadblocks and Interdisciplinary Challenge Team Data Gaps

The ICT screening process, which involved comparing questionnaire responses and reviewing data management plans, revealed seven recurring obstacles that limit the efficient flow of genomic data from project teams to downstream users. These challenges are technical, organizational, and cultural in nature, and collectively may hinder progress toward achieving full interoperability and data reuse, ultimately limiting the long-term impact on GHG emission reductions in the agricultural sector. The roadblocks are as follows:

1. **Fragmented Data Storage and Inconsistent Transfer Methods**

   Many ICTs rely on local storage, portable hard drives, or short-term cloud links for data transfer. This fragmentation creates risks of data loss, version confusion, and inequitable access among partners. Moreover, these practices expose research genomic data to varying degrees of cybersecurity risk.

   Recommendation: Adopt a shared, persistent repository infrastructure with controlled access, integrated with a harmonized metadata catalogue. Use APIs and secure data exchange protocols to automate submission and retrieval. Encourage cybersecurity training for the project team.

2. **Uneven Metadata Documentation**

   Metadata preparation varies widely, with incomplete sample identifiers and missing contextual details reducing long-term usability.

   Recommendation: Implement homogeneous metadata templates aligned with MIxS and the Crop Ontology, along with metadata validation prior to any data submission to national or international repositories as well as manuscript publication. Automated metadata checkers and training for data managers can help ensure uniform quality.

3. **Limited Incentives and Cultural Barriers to Data Sharing**

   Researchers hesitate to deposit data due to concerns over credit, ownership, and additional workload.

   Recommendation: Introduce clear data ownership policies, the use of persistent identifiers (DOIs and ORCIDs), and institutional recognition for data contributions beyond a project's lifecycle. Provide user-friendly submission tools and clear attribution pathways to encourage participation.

4. **Inconsistent Governance and Access Control**

   Data access levels and permissions settings differ across projects, sometimes delaying analyses and sharing, negatively impacting data reuse.

   Recommendations: Develop a unified data governance and management framework describing access tiers, consent procedures, and retention periods, consistent with federal and provincial regulations. This framework must be harmonized with international scientific community standards, such as the Global Alliance for Genomics and Health and the FAIR principles. Acknowledging data ownership and team members' contributions to data generation should also be considered in data governance.

5. **Insufficient long-term preservation and sustainability planning**

   When funding ends, many datasets risk becoming only internally useful or inaccessible due to lack of maintenance and archiving resources.

   Recommendation: Create a strong long-term data stewardship plan that designates a repository of records and specifies responsibilities for curation and migration. National infrastructure partners should provide persistent hosting and archival storage.

6. **Training and Capacity Gaps**

   Uneven technical skills in bioinformatics and data management lead to variable data quality and pipeline reproducibility.

   Recommendation: Develop and promote coordinated training modules on data management, metadata harmonization, workflow reproducibility, and FAIR data principles. Encourage cross-project mentorship and workshops to build collective expertise. Encourage initiatives that foster the education of the upcoming generation of researchers, emphasizing a contemporary culture around data practices, collaboration, and sharing of information.

7. **Policy and Interoperability Gaps**

   Different provincial and territorial rules for data privacy, access, and environmental reporting create additional friction.

   Recommendation: Adopt interoperable metadata frameworks to bridge jurisdictional differences and support a national agri-genomics data strategy.

3.6. Recommendations for Interdisciplinary Challenge Team Data Tools

The recommendations for data tools presented in this section are intended to effectively meet both present and anticipated challenges related to the generation, use, and sharing of agri-genomic data. It is important to adopt coordinated strategies that combine common data tools, governance frameworks, and long-term planning. The recommendations synthesize findings from the literature review, the international interviews, the ICT information collection matrix, and the Ag Data Commons (see Appendix C).

3.6.1. Implement Cross-Project Governance and Coordination

Goal: Establish a clear, shared governance structure that maintains consistent practices across ICTs while also supporting project-specific flexibility.

The shared governance structure will serve as a forum for approving templates, operating procedures, catalogue rules, and training plans. Shared principles should reflect the FAIR principles for technical interoperability and the CARE principles for Indigenous

data governance. In addition, it should identify when the Nagoya Protocol access and benefit-sharing obligations apply [84].

Recommended actions:

- **Establish a cross-project data committee.** The committee should include representatives from all ICT projects and research funding agencies, and should have a written charter, decision rules, and quarterly meetings to propose and implement common data governance, management, storage, and sharing rules.
- **Define shared principles for access, ownership, reuse, and retention.** Principles should align with federal and provincial policies and program data sharing requirements. Exceptions for sensitive or proprietary datasets should be documented.
- **Approve and version shared artifacts.** A single register should be maintained for approved metadata templates, operating procedures, quality control documentations, file naming rules, and catalogue submission guidelines. Semantic versioning and publish change logs should be used.
- **Set review and audit cycles.** Scheduled reviews of data management practices, metadata completeness, quality control conformance, and catalogue records should be established. Corrective actions should be tracked to closure.
- **Resolve cross-project issues.** A simple system should be used to log items that prevent data integration from running smoothly, along with general questions and repository issues.

Impact and simple metrics: This shared governance structure will reduce ambiguity, shorten time to compliant release, and improve trust in shared outputs. Tracking should include the number of approved artifacts, time from draft to approval, first-pass compliance rates for submissions, the count and resolution time of cross-project issues, and attendance and participation in working group meetings. Practices should be adjusted on a regular basis (e.g., quarterly) based on these indicators.

3.6.2. Develop and Adopt Unified Metadata Templates

Goal: Make agri-genomic datasets consistently understandable and reusable across ICT projects and institutions by using shared, field-tested metadata templates that align to recognized scientific community standards.

Recommended actions:

- **Anchor to the right community standards.** MIxS should be used for sequence and environment context, MINSEQE for sequencing experiment description, and MIQE guidelines for quantitative and digital PCR. Archive and identifier practices should be aligned with GenBank and BioProject submission schemas, BioSample identifiers, dataset DOIs, and contributor's ORCID. Core formats should be homogenized to FASTQ for raw reads, BAM or CRAM for alignments, and VCF for variants, with explicit reference to genome build identifiers. Where applicable, metadata elements and data use terms should be aligned with Global Alliance for Genomics and Health models.
- **Publish project-wide templates by data type.** Concise JSON or spreadsheet templates should be provided for amplicon sequencing, metagenomics, whole genome sequencing, array genotyping, RNA-seq, and digital PCR; and should include field names, definitions, allowed values, and examples.
- **Define required and recommended fields**. Required metadata fields should cover sample origin and standard identifiers, collection date and location, platform and kit used, library and run parameters, reference genome, quality control metrics, data processing steps, repository accession identifiers, and related publication identifiers. Recommended fields should cover controlled vocabulary terms and links to protocols.
- **Provide controlled vocabulary.** Consensus-based standard terminology for agri-genomic data should be adopted. Accepted ontologies for organism, tissue, environment, and traits should be referenced, and short lookups for common terms should be supplied to reduce free text.

- **Pilot, refine, and adopt.** Short pilots should be run in each project, comments from data managers and analysts should be gathered, and templates should be revised before adopting the final set for all new datasets with accurate documentation. Legacy backfill should remain optional but encouraged.

Impact and simple metrics: Shared metadata templates will reduce submission friction, improve machine readability, and accelerate cross-project synthesis. Tracking should include the proportion of new datasets that use the provided templates, the average number of missing required fields per submission, the time from first upload to acceptance, and the number of successful cross-project dataset joins that rely on template fields. Metadata templates should be adjusted on a regular basis (e.g., quarterly) based on these indicators.

3.6.3. Harmonize Data Processing and Quality Control

Goal: Make data pre-processing and quality control consistent so results are reproducible and comparable across ICTs.

Recommended actions:

- **Develop shared operating procedures for each data type.** As part of the cross-project data committee proposed in Section 3.6.1, procedures should be established and maintained specific to all types of genomic datasets, such as amplicon, metagenome, whole genome sequencing, array genotyping, RNA-seq, and digital PCR. Recommendations from the Genomic Standards Consortium and AgBioData guidance should be drawn on to define the operating procedures required for each data type. Required steps, necessary tools or acceptable alternatives, input and output formats, and parameter ranges should be clearly specified.
- **Define minimum quality control and reporting.** For sequencing: adapter and quality trimming, contamination checks, alignment settings, duplicate handling, and coverage targets. For variants: caller settings, site and sample filters, reference build, and lift over rules. For arrays: call rate, concordance, and Hardy–Weinberg checks as appropriate. For RNA-seq: strandedness checks, mapping rate thresholds, gene model version, and count normalization method. A short, machine-readable quality control summary file (e.g., JSON, TSV) that captures key metrics should be published so automated validation is possible without changing existing analysis tools.
- **Pin references and versions.** Explicit reference genome identifiers, gene model versions, ontology releases, and containerized tool versions should be required. These steps should be documented in the quality control summary and workflow metadata to ensure consistent record-keeping and facilitate effective data reuse.
- **Provide runnable workflows.** Containerized pipelines and notebooks should be provided for each operating procedure, with examples of data inputs and expected outputs. These should be hosted in a shared repository and stable releases should be archived.
- **Pilot, revise, and adopt.** Operating procedures should be tested on representative datasets from each project, feedback from data managers and analysts should be gathered, and procedures should be revised before adopting them as the program default for new datasets. Justified deviations should be allowed with a short variance form that records what changed and why.

Impact and simple metrics: Shared operating procedures will raise data quality, shorten repository submissions, and reduce rework. Tracking should include acceptance rates on first submission, average time from upload to approval, number of variance forms filed, concordance rates on replicate analyses, and the proportion of datasets that include a valid quality control summary file. These indicators should be reviewed on a regular basis (e.g., quarterly) and operating procedures updated as needed.

3.6.4. Create a Shared Discovery Catalogue and Repository Policy

Goal: Give every ICT project dataset a durable home for discovery while letting files live in the most appropriate domain or general repository.

Recommended actions:

- **Adopt a repository-plus-catalogue model.** ICTs should deposit data in repositories (e.g., NCBI Sequence Read Archive, Gene Expression Omnibus, European Nucleotide Archive, Zenodo, or MG-RAST), then each deposit should be registered in a Canadian repository (e.g., Digital Research Alliance of Canada or Genome Canada catalogue) that serves as the authoritative index for the program.
- **Require a persistent record per dataset.** Each catalogue entry should have a DOI landing page, machine-readable metadata aligned to DCAT or DCAT-US core fields, direct links to the source files, and bidirectional links to related publications. ORCID should be supported for contributor identity.
- **Respect access controls.** The catalogue should only expose non-sensitive discovery metadata. Authorization and file access should remain with the source repository for controlled datasets.
- **Enable automation.** API endpoints should be provided for harvesting and reuse. The quality control documentation file produced under the shared operating procedures should be implemented so automated checks can run at ingest.
- **Pilot, then mandate.** Projects should start with one dataset to validate field mappings from the unified templates and to test automated metadata harvesting. After the pilot, catalogue registration should be required for all new datasets and backfilling of legacy outputs should be scheduled.
- **Operate for longevity.** Routine backups, checksum verification, and metadata versioning should be implemented. A clear deprecation and redirection policy should be published if sources relocate. Ad hoc hard drive transfers should be phased out in favour of secure managed transfer endpoints or object storage, with checksum verification and lifecycle rules. ICTs should be aware of cybersecurity risks associated with data sharing practices and genomic data longevity.

Impact and simple metrics: A shared catalogue will improve findability and stewardship without centralizing every file. Tracking should include the data sharing practices with a valid catalogue record, the time from dataset release to catalogue registration, API usage for metadata retrieval, the rate of broken links detected and repaired, and the proportion of records with ORCID, DOI, and publication links. These indicators should be reviewed on a regular basis (e.g., quarterly) and catalogue policies refined as needed.

3.6.5. Enable Cross-Project Data Integration and Reuse

Goal: Facilitate the integration, comparison, and reuse of datasets produced by the ICTs to ensure cross-project synthesis becomes common practice.

Recommended actions:

- **Set minimum integration specifications.** Unified metadata templates and shared operating procedures should be used to define required fields and formats for key data types. For example, linking genotypes with sequence data should use consistent sample identifiers, metagenome comparisons should rely on MIxS fields and agreed environmental terms, RNA-seq matrices should use harmonized gene and transcript identifiers, and variant data should be provided in VCF format with the reference build and applied filters clearly declared.
- **Harmonize naming and versions.** Common file naming conventions, reference genome versions, and core formats should be adopted across projects. Shared identifiers for samples, individuals, sites, and publications should be enforced.
- **Provide practical utilities.** Lightweight tools for validating field definitions, renaming files, and mapping identifiers should be provided. Simple

visual summaries should be included, such as quality control dashboards, principal component analysis plots for genotype and expression data, and interactive abundance profiles for metagenomes.

- **Ship maintained code and examples.** Short workshops or code sprints should be run to co-develop and test these tools on real ICT datasets. Versioned, containerized scripts and notebooks should be published in a shared GitHub organization with clear README files, example inputs, and test data. Each ICT should be required to contribute at least one reproducible example pipeline that consumes another ICT's public data.
- **Expose programmatic access.** Stable APIs that return machine-readable catalogue metadata consistent with the catalogue approach in Appendix C should be provided, which will facilitate scripted cross-project discovery and retrieval.
- **Document the playbook.** Integration practices should be documented in a concise guide that covers identifiers, formats, minimal fields, and validation steps.

Impact and simple metrics: Cross-project data integration will reduce friction at merge points and facilitate more comparable analyses and routine cross-project synthesis. Tracking should include the proportion of datasets that pass integration validation, the number of cross-project pipelines executed per quarter, the time required to integrate a new dataset, and reuse events measured by API calls, GitHub downloads, and citations of example pipelines.

3.6.6. Support Training and Capacity Building

Goal: Ensure the sustained use of shared scientific community standards and tools by investing in the people who apply them. This supports the recommendations in sections 3.6.1 through 3.6.4 by building skills in data governance and management, metadata templates, operating procedures, pre-processing and quality control, and catalogue registration.

Recommended actions:

- **Establish a continuous learning program.** Recurring workshops and webinars should be offered, along with on-demand modules on topics such as using the unified metadata templates, completing operating procedures checklists, submitting to NCBI and other repositories, data pre-processing, quality control and troubleshooting, and preparing DCAT-aligned catalogue records. Recordings, slides, and example files should be archived in a central library.
- **Provide dedicated user support.** A helpdesk with office hours and a ticketing system should be established, along with a living FAQ that captures resolved issues, links to code examples, and points to the correct templates.
- **Develop a peer mentorship network.** A data lead, who serves as a local expert and data practices trainer, should be identified for each project. A train-the-trainer model should be used to spread skills quickly and create a community of practice across agri-genomic projects.
- **Create role-specific resources.** Onboarding checklists for new personnel, curator playbooks for metadata review, analyst playbooks for quality control and pipeline runs, and principal investigator compliance checklists for release timing should be published. Small sandbox datasets for amplicon, metagenome, whole genome sequencing, array, RNA-seq, and digital PCR should be included so users can practice end to end.

Impact and simple metrics: This training and capacity building structure will reduce barriers to adoption, improve data quality, and keep research moving. Tracking should include completion of core modules, first-response time on helpdesk tickets, the number of submissions that pass automated checks on first attempt, and the number of cross-project datasets successfully integrated. Topics should be adjusted on a regular basis (e.g., quarterly) based on these indicators.

3.6.7. Monitor, Evaluate, and Refine Implementation

Goal: Maintain ICT program effectiveness by regularly evaluating outcomes, addressing inefficiencies, and updating scientific community standards and tools as required.

Recommended actions:

- **Create a review schedule.** Short monthly check-ins should be held for data leads to surface issues, and a deeper quarterly review should be led by the cross-project data standards working group to assess compliance, blockers, and priorities.
- **Track clear indicators.** Clear indicators should be tracked by maintaining a shared dashboard with the following set of program metrics:
  - Adoption (share of new datasets using the unified templates and operating procedures);
  - Quality (pass rates for quality control summaries and metadata validation);
  - Reuse (cross-project downloads, API calls, and citations of example pipelines);
  - Timeliness (median time from data generation to a registered catalogue record); and
  - Coverage (percent of legacy datasets backfilled into the catalogue).
- **Run lightweight audits.** On a regular basis (e.g., quarterly), a small number of datasets should be sampled from every project to verify required fields, identifier consistency, declared reference builds, and presence of quality control summaries. Written feedback should be provided within two weeks.
- **Collect structured feedback.** A single feedback form and a ticket queue should be used for issues with templates, operating procedures, or the catalogue, and themes should be summarized ahead of each review.
- **Manage change with versioning.** Semantic versioning should be applied to templates, operating procedures, and APIs. A brief changelog, a deprecation policy with realistic timelines, and migration notes should be published. Breaking changes without a pilot and documented rollback should be avoided.



- **Close the loop.** After each regular review (e.g., quarterly), a one-page action log that lists decisions, owners, and due dates should be posted. Open items should be revisited at the next meeting.

Impact and simple metrics: Ongoing monitoring, evaluating, and refining will result in continuous improvement without disruption, higher data quality, and faster reuse. A certain compliance percentage should be targeted for new datasets with templates and operating procedures.

3.6.8. Secure Funding for Long-Term Data Preservation

Goal: Ensure the continued accessibility, integrity, and usability of ICT datasets beyond individual project timelines by establishing sustainable funding and governance for long-term preservation.

Recommended actions:

- **Establish dedicated budgets.** A dedicated budget for data archiving and stewardship should be established that is separate from short-term research funds.
- **Create concise preservation plans.** Each project should be required to submit a concise preservation plan specifying the datasets to retain, the responsible parties, repository locations, retention timelines, and verification procedures.

- **Use a hybrid model.** Data should be deposited in trusted domains or general repositories (e.g., NCBI, European Nucleotide Archive, Zenodo) and a central catalogue should be maintained for discovery and long-term tracking.
- **Establish partnerships with repositories.** Program-level agreements should be negotiated with national and international repositories to secure storage quotas, ensure metadata interoperability, and guarantee long-term access.
- **Create clear and detailed budgets.** Budgets should explicitly include storage, periodic integrity checks, metadata curation, and backfilling of legacy datasets.
- **Create a review schedule.** A lightweight review cycle should be implemented to monitor dataset availability, persistent identifiers, and link validity.

Impact and simple metrics: By ensuring long-term funding and repository partnerships, ICTs can maintain genomic data as durable public assets rather than ephemeral project outputs, which will safeguard investments, support reproducible science, and provide ongoing access for regulators, policy-makers, industry, and the broader research community.

3.7. Workshop Validation and Interviews

Throughout the workshop, participants engaged in discussions by responding to a series of predetermined questions designed to assess their backgrounds, including their areas of expertise (Figure 1), their experience with genomic data (Figure 2), and areas where they felt data standards may or may not be required (Figure 3 and Figure 4). In addition, participants were asked whether a metadata template was available for use by all of their team members working with genomic data.

Workshop participants came from diverse professional backgrounds, including knowledge mobilization, animal scientists, geneticists, and those with laboratory management experience. In addition, most of the participants worked with array data and amplicon sequencing. When participants were asked about the availability and use of metadata templates across team members, 50% indicated that they do not have a metadata template in place that is shared and used within their projects. Not only does this highlight the need for harmonization of data practices among genomic data users, but it also highlights the importance of implementing standards for metadata preparation

Figure 1: Areas of expertise and professional focus of workshop participants.

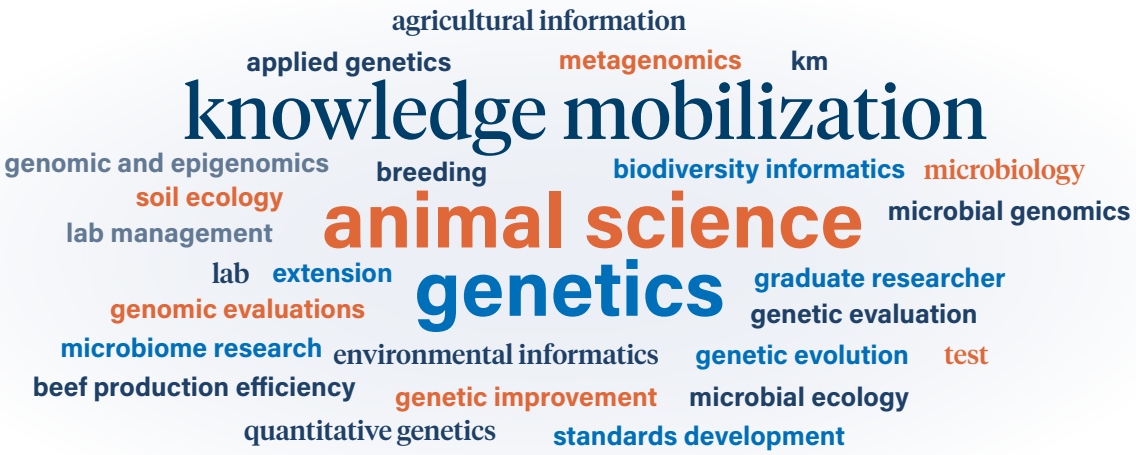


Figure 2: Types of genomic data managed by workshop participants.



with minimum requirements and controlled vocabularies to support and leverage data storage and reuse.

As noted by the majority of participants, the adoption of standards is critical for ensuring consistent metadata documentation within and across projects, when working with phenotypic and genomic data, and for publication of research results. Additionally, most of participants indicated that standards are needed throughout the entire process, from genomic data generation through analysis to application or publication of results.

Although participants indicated that standards are not needed for data analysis and innovation (from hypothesis to model definitions), or for results interpretation, most of them also indicated that standards are not applicable for scientific publications, which is contradictory because most international journals require that data underlying publications be made publicly available and accessible, in accordance with journal-specific quality standards and requirements, prior to publication (e.g., [18], [107], [108]). In other words, agri-genomic researchers are expected to adhere to certain international quality standards when preparing their results and data for publication.

The final three questions were discussed after presenting the eight recommendations for genomic data practices outlined in Section 3.6. The goals of these questions were (a) to understand the participants' perspectives in terms of the value and ease of implementation of the recommendations, (b) to rank the recommendations from most important to least important, and (c) to evaluate whether the participants felt that any recommendations were missing and should be incorporated into the proposed list.

As shown in Figure 5, participants ranked five of the recommendations as highly valuable but also difficult to implement, with particular emphasis on implementing cross-project governance and coordination, and securing funding for long-term data preservation. Participants also ranked two recommendations as less difficult to implement but highly valuable for genomic data standardization: developing and adapting unified metadata templates, and supporting training and capacity building. Interestingly, even though they considered it somewhat easy to implement and highly valuable for research purposes, 50% of the participants indicated they did not have a metadata template in place, highlighting the importance of harmonized practices that incorporate the consistent use of proper genomic data practices and documentations (Figure 6).

Figure 3: Areas where workshop participants identified a need for standards.



Figure 4: Areas where workshop participants identified that standards are not needed.



Figure 5: Workshop participant ranking of proposed recommendations according to perceived value and difficulty of implementation.

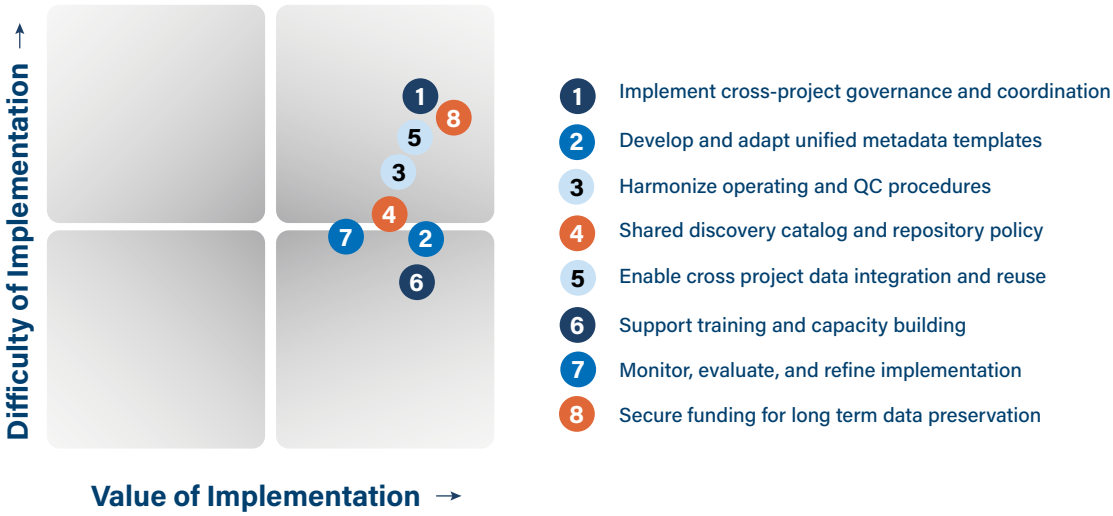
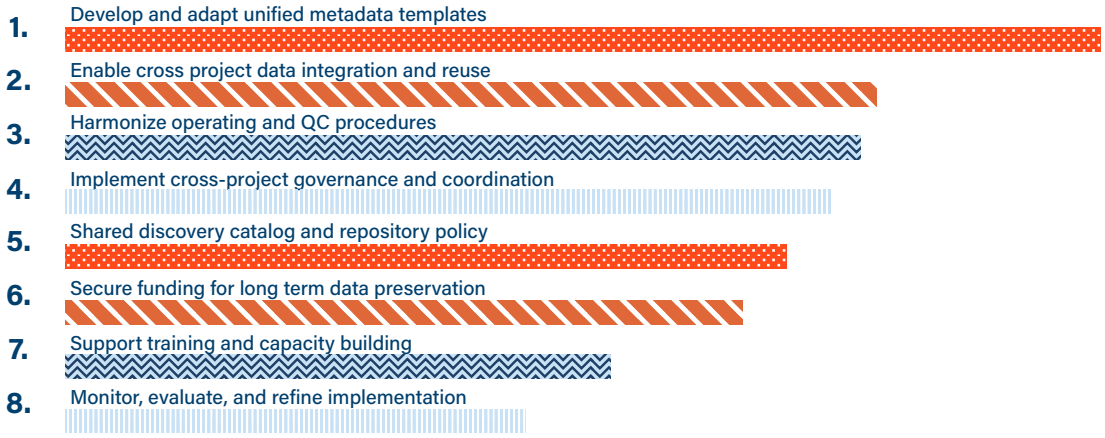


Figure 6: Ranking of proposed recommendations based on their perceived importance.



The responses to the second question were consistent with the responses concerning the prioritization of recommendations based on importance. As shown in Figure 6, workshop participants ranked the development and adoption of unified metadata templates as the highest importance, followed by facilitating cross-project data integration and reuse, and harmonizing operational and quality control procedures, including comprehensive quality control documentation for data quality assurance.

Finally, participants were asked whether any recommendations should be added to the list, aiming to identify any aspects that were not captured during the literature review, interviews, and ICT document screening. Although one participant indicated that a recommendation was missing, no further information was provided or suggested throughout the workshop. Based on follow-up discussions with workshop participants, it was determined that the recommendations are comprehensive, and no additional recommendations are required to start implementing agri-genomic data standards in Canada.

The presentation and interactive question portion of the workshop was followed by a breakout rooms session. The following questions were provided to guide the group discussions:

1. Identify and discuss situations where standards on genomic data are applicable.
2. Discuss the relevance of data standards to:
    a. data users (researchers and laboratories);
    b. end users (industries);
    c. consumers; and
    d. policy-makers.
3. Based on your experience, are there any recommendations missing that should be considered? If so, list them.
4. Who are the group(s) or party(ies) that would be directly impacted by implementing data standards? Why?

A general comment that emerged from the breakout rooms was that it is important for the Canadian agriculture sector to have genomic data practice standards in place, not only to leverage data quality and reuse but also to avoid relying on international repositories for storing data. Addressing the need for data harmonization and standards would also provide research funders with a structured and consistent approach to request the standardization of genomic data practices when investing in research projects. Participants emphasized that it would be more effective if standards for data governance and practices were promoted and enforced by funding agencies because funders would have the required strength to implement data standards, metadata

templates with the minimum information required, as well as proper standards for data sharing. Furthermore, policy and governance differences among funding institutions would be minimized by the implementation of similar practices across funders. In other words, data practices, sharing policies, and standards could be promoted by a consortium of research funding agencies across Canada, which corroborates and refines the proposed recommendation to implement cross-project governance and coordination.

The breakout rooms discussions also highlighted the need for further consideration of data ownership, which should be integrated into the harmonization of agri-genomic data practices. This is particularly challenging in co-funded public–private projects, when the industry supports researchers with phenotypic and genomic data collection. It was mentioned that because the industry primarily relies on its data to implement new business directions, short- and long-term data ownership should be part of research project data governance standards. Furthermore, considerations should include how industry data is shared and the potential implementation of reuse fees on existing datasets. These considerations emphasize the important role of research funding institutions in promoting standardized practices around agri-genomic data in Canada. Another issue that was highlighted during the workshop is that genomic information is meaningless without phenotypic records and corresponding metadata. Thus, proper documentation of research project design, data collection, and phenotypic data is essential to complement genomic information. All this information should be incorporated into the metadata with a standardized minimum requirement that leverages the use and reuse of such data. If metadata and documentation follow standard procedures, data reuse could reduce the need for new data generation and save project funds.

The group discussion also revealed that there are no missing recommendations and that standards are essential for data pre-processing and quality control, supported by comprehensive documentation of all procedures undertaken. However, standards do not apply to statistical analysis, as each project will have its own objectives. In general, participants indicated that the proposed recommendations are a good start to promote the implementation of standards for genomic data practices in the Canadian agricultural sector. They also indicated that the most impacted parties will be data users (researchers and laboratories) and industries (the end users of research outcomes), but consumers will be indirectly affected because knowing that the data was collected and managed following Canadian standards might give them more confidence in the research results. Thus, by having well-established harmonization and standards in place, the agricultural sector would benefit from a clear understanding of all practices that were implemented, from data collection through application, enhancing the clarity of the results among different interested parties. The workshop concluded with participants' reflections from the breakout rooms, focusing on the applicability and challenges of implementing data standards. Participants highlighted the importance of metadata harmonization as a top priority for enabling interoperability and long-term data reuse. The closing discussion also emphasized the need for clear governance framework, including funder-driven polices. Training and capacity building were also identified as critical for adoption, alongside incentives for researchers to share their project data. Finally, data users (researchers and laboratories) and end users (industry) were identified as the parties that were likely to be most impacted by the introduction and implementation of new data tools, including standards.

To expand and validate the workshop discussions, follow-up interviews were conducted with four representatives from the two parties identified as potentially most impacted by the introduction and implementation of new genomic data tools. These semi-structured interviews were guided by a brief questionnaire focusing on three key areas that workshop participants indicated as requiring further attention for promoting genomic data harmonization and standardization in Canada (see Appendix D). To address the different roles of parties affected by genomic data standardization, the interviews included participants with different backgrounds in the agri-genomic data field, including a data standardization expert, a research data user, an industry consultant with consortium experience, and an industry end user. The questionnaire



incorporated aspects related to cross-project data governance and management with consistent use of harmonized metadata templates, the need to establish and promote unified Canadian standards for agri-genomic data, and the importance of data ownership and sovereignty over both agri-genomic information and data repositories.

3.7.1. Use of Metadata Templates

During the semi-structured follow-up interviews, interviewees highlighted that the implementation and use of metadata is a primary barrier around agri-genomic datasets within and across different projects and parties, influencing data consistency and reuse. Interviewees described how missing or inconsistent identifiers, unclear headers or definitions (from proper identification of sampling location to inconsistent information about data processing), and incomplete contextual information force downstream users to return to the originating team to interpret what the data represents and whether it is fit for a new research purpose. Specifically, one of the interviewees emphasized that metadata templates are effective only when consistent standards and vocabulary are used by all project members. According to the same interviewee, metadata templates should not be project-specific because sometimes the same fields may have different interpretations depending on the group using them. As suggested in this report, metadata templates should be genomic datatype specific, incorporating the MIxS approach to provide the minimum information required. This was corroborated by another interviewee, who emphasized the need for well-defined and stable sample identifiers (e.g., use of standard animal registration numbers or identification) and naming conventions to enable records to be linked across multiple systems with minimal back-and-forth beyond the project lifecycle. It was also pointed out that the quality of metadata tends to vary considerably across groups, which can cause a pragmatic bottleneck to interoperability. The general theme that emerged from the follow-up interviews validated the workshop discussion: a baseline standardized metadata template with a common vocabulary across projects would serve as the foundation for cross-project governance and data reuse.

3.7.2. Cross-Project Data Governance

The follow-up interviews also corroborated observations and discussions from the workshop around standardized data governance, reinforcing that genomic data sharing and reuse depends on both effective governance and technical compatibility. Interestingly, data governance emerged as a practical

set of questions, such as who approves access, what conditions apply, how requests are processed, and who acts as steward as projects evolve. Several different perspectives were highlighted by the interviewees, including:

- Data governance and management practices are effective levers when expectations are determined early and treated as measurable outputs rather than just an optional guidance.
- Cross-project collaboration within industry or consortia settings often relies on clearly defined scoping and approval processes from both parties, since unrestricted sharing is rarely feasible from a commercial point of view.
- In practice, reuse constraints are frequently reduced to explicit terms, permissions, and handling requirements captured in data sharing agreements, which most of the time is not a common document.

Taking all the previously discussed aspects into account, these findings support the need to strengthen shared governance mechanisms, such as consistent approval workflows, clearly defined stewardship roles, and standardized agreement terms, alongside any technical standardization efforts.

3.7.3. Data Ownership and Sovereignty

Concerns about data ownership and sovereignty that were pointed out during the workshop were corroborated by the follow-up interviews. The four interviewees emphasized that genomic data ownership and sovereignty can delay or in some cases impede data sharing and reuse beyond the original scope of a project, even when the datasets are technically usable and curated for sharing. Practical concerns included several questions, such as who owns a dataset, what downstream analyses and uses are allowed, and what risks are tied to data release. These factors also impose limitations on manuscript publications and how agri-genomic data can be reused across different projects prior to and following a project's completion. Another important aspect related to ownership and sovereignty is the safeguard of confidential data and sensitive information. One potential strategy for this is the implementation of a standardized method for editing private details. This approach should use metadata to protect and link original identifiers to sharable identifiers and content, enabling controlled access while maintaining data privacy when necessary. One of the interviewees recalled a situation where ambiguity around ownership and sharing risks impacted publication and noted that redaction may be required before some genomic information can be openly shared.

Another interviewee, with a consortium/industry background, shared a different perspective, stating that competitive constraints often necessitate preliminary permission-based access and approval processes rather than open deposit in a data repository. Moreover, the end user (industry) participant noted that ownership expectations boil down to "if you paid for it, it's yours," and that multi-party co-funding raises practical questions about access and permitted uses. The discussion around data ownership also emphasized the importance of common data sharing agreement terms, in which long-term data ownership is clearly acknowledged. According to one of the interviewees, the role of formal data sharing agreements is to define responsibilities and permitted uses, including post-project obligations (e.g., return or destruction of datasets), and in some cases, jurisdictional constraints requiring that data remain in Canada unless reapproved. The agreement terms should be defined by the cross-project data governance team, including minimum requirements that enhance the legacy of agri-genomic data in Canada. All these considerations are important to support the research findings and the workshop discussions, highlighting the need for explicit data ownership rules and clear data sharing agreement structures, allowing risk-free data sharing practices and reuse in all types of agri-genomic research projects (and particularly relevant in co-funded and multi-party settings).

Given the nature of semi-structured interviews, participants discussed additional elements of data standardization that extended beyond the questionnaire's three focus areas. It was noted that data standards and metadata templates must be clear

about which level of data they apply to, distinguishing raw data holdings from project-specific processed datasets and publication-oriented deposits, as each data level implies different expectations and practical requirements for sharing and reuse. One interviewee pointed out that pre-processing and quality control procedures must be standardized internally and cannot be treated as an afterthought. This requires proper documentation because the datasets are frequently shared in edited or transformed formats and, in some settings, raw signals may be unrecoverable when collection platforms apply embedded algorithms or workflows that change what is retained. This aligns with the broader understanding that transparent documentation of data processing steps and quality control supports trust among parties and enhances reuse alongside the results themselves.

According to one interviewee, the implementation of good data practices succeeds when expectations are set early (prior to starting a project) and supported by practical infrastructure and expertise, so data governance and management requirements are feasible rather than an added burden. Another participant questioned the value of fully centralizing all data into a single integrated database, and instead described approaches that prioritize discoverability and governed data access. The follow-up interviews provided qualitative support for the perspective that the adoption of data practices toward standardization improves when proposed standards are paired with sustained resourcing and operational support (from institutions and funding agencies), not solely with the publication of templates at the outset.

3.7.4. Refine Standardization Needs

During the four follow-up interviews, participants contributed insights that informed the refinement of the findings outlined in this report. For example, interviewees clarified what metadata standardization means operationally, indicating that templates are only effective to improve interoperability when users can complete them using predetermined, controlled vocabulary that is automatically presented to them (i.e., consistent definitions and constrained values rather than free-text fields). One interviewee emphasized that a practical approach requires prioritizing consistent stable identifiers as core join keys but cautioned against making templates too detailed such that they become prescriptive across diverse and evolving data types. This was corroborated by another participant, who reinforced the concern in the context of data legacy, describing how phenotypes and genomic information collected across distinct projects and laboratories often differ in names, units, and formats, making data consolidation "not a small job, but … doable" with dedicated effort to produce mergeable datasets. Thus, standardization of data practices should be framed as a mandatory metadata core (join keys and minimum required context) with controlled flexibility for project- or data-specific variation, rather than assuming a single universal template will apply in all cases.

Regarding cross-project data governance and management, the follow-up interviews sharpened governance from a general principle into an explicit operating requirement. Interviewees emphasized that data governance becomes practical when it is linked to funding agencies. In other words, data governance and management practices should be established at the outset and treated as measurable deliverables rather than optional documentation. One interviewee noted that barriers are not only technical, but that social and institutional dynamics can also block sharing, even when technical obstacles are not the primary issue associated with genomic data. This implies that governance needs a workable decision-making process for access, sharing, and reuse, including clear stewardship responsibilities as projects evolve and beyond the project lifecycle. Given that cross-institutional sharing is often constrained by intellectual property considerations and conditions set by participating parties, data governance must operate within real constraints rather than assuming open sharing by default. Thus, as an overall refinement from the workshop discussions, data governance and management practices should be framed as an important resource and accountable operating model (e.g., decision rights, processes, and stewardship), not merely as a policy statement.

Two refinements proposed in the follow-up interviews do not fall clearly within the three areas that guided the interviews (see Appendix D), but they contribute important layers to how standardization should be framed in the agri-genomic space. The first is to distinguish standards for future data collection from those required to consolidate historical datasets. Interviewees emphasized that legacy harmonization requires sustained manpower and if they are not actively stewarded, datasets can effectively disappear when researchers retire. The second refinement is that standards must connect to infrastructure. According to the interviewees, in operational terms, this becomes a set of concrete questions: Where does the metadata live, who maintains it, and how does it remain accessible and usable over time? Overall, these refinements are pragmatic and do not reverse the workshop's findings; instead, they translate standardization into implementable components (e.g., a baseline metadata core, feasible data governance pathways, agreement-backed sharing terms, and resourced stewardship) that can be evaluated and maintained.

Finally, the follow-up interviews largely validated the eight recommendations presented in Section 3.6, but they also highlighted three elements that were not explicitly presented as recommendations. First, interviewees sharpened the need to treat data tiers and data products appropriately. In other words, the standardization process should specify whether it targets raw genomic datasets, project-level pre-processed datasets, or publication deposits. This is important because each tier carries different expectations, working hours, costs, and documentation needs (in some cases, raw signals cannot be reconstructed once platforms and workflows transform what is stored). Second, interviewees added a clearer recommendation to formalize standard data agreement rules, including ownership definitions across the data lifecycle, approval-based access options, and sovereignty clauses, such as jurisdictional restrictions and end-of-project return or destruction unless reapproved. Third, interviewees emphasized that standards need an explicit maintenance path that includes a named steward or owner, a place where



metadata and supporting documentation live, and a simple change process so standards do not degrade as momentum fades.

4. Recommendations for Agricultural Genomics and Conclusions

As part of the Genome Canada Climate Action initiative, which is aligned with the Canadian Net-Zero Emissions Accountability Act, the agricultural research sector is generating high-value genomic datasets that are expected to deliver significant climate mitigation benefits in the coming years. By combining researchers, business, farmers and communities, the ICTs are focusing their efforts on reducing GHG emissions and the environmental impact of Canada's agricultural practices. Through the screening of data practices across ICTs (from generation to application), combined with insights from internationally recognized initiatives, a review of relevant literature, workshop discussions, and interviews with industry parties who will be directly affected by the implementation of new data tools, it is clear that there is an opportunity for Canada to advance toward coordinated, structured,

and sustainable national data standards, while matching international scientific community standards.

At the time of writing, despite the progress toward and the benefits expected from reducing Canada's agriculture GHG emissions, the long-term impact of investments in agri-genomic research is limited by inconsistent data practices, fragmented data governance and management, and the lack of or use of inconsistent metadata documentation (from data collection and pre-processing to quality control and sharing processed data). These factors highlight the opportunity for harmonization of data practices, which would enhance data interoperability, data sharing, and data reuse. To address these limitations, this report proposed a set of recommendations, outlining a practical, actionable framework to improve data quality and cross-project collaboration while supporting the legacy of agri-genomic research in Canada. The eight recommendations are summarized in this section, along with details about the relevant parties that would benefit from the implementation of each recommendation.

Recommendation 1: Implement Cross-Project Governance and Coordination

It is recommended to establish a formal governance structure, led by a cross-project governance committee, that includes representatives from different projects, funding agencies, and data specialists. This committee would collaborate to ensure consistent governance and data practices across agri-genomics in Canada. The committee could also determine data ownership rules, access tiers, data retention periods, levels of data sensitivity (for both research and commercial purposes), and promote the benefits of data sharing and reuse. In addition, the committee could provide a stable governance framework and conduct regular review cycles to reduce ambiguity, strengthen accountability, and improve transparency among various parties, such as researchers, funding organizations, and end users.

All relevant parties (academia, industry, commercial laboratories, and industry–academia groups) will be influenced by this recommendation, as the cross-project governance committee should include representatives from each sector, working collaboratively to establish common data practices. Furthermore, should commercial laboratories and industry parties already have data coordination groups in place, they can still benefit by adopting recommendations from the committee into their practices.

Recommendation 2: Develop and Adopt Unified Metadata Templates

It is recommended to promote standardized metadata templates that are aligned with scientific community standards. These templates should encompass project design, environmental context, and MIxS (where applicable). Metadata templates could be provided with clearly defined required fields, controlled vocabularies, and structured formats, including stable and consistent sample identifiers. In addition, the use of metadata templates tailored to each primary genomic data type could be considered as part of standardized practices.

The adoption of unified metadata templates is expected to benefit all relevant parties working with genomic data (academia, industry, commercial laboratories, and industry–academia groups), enhancing data usability and supporting long-term integration.

Recommendation 3: Establish Standardized Data Processing and Quality Control

It is recommended to establish standardized pre-processing and quality control protocols, along with documentation. This will minimize the negative effects of evolving genomic platforms, variations in processing pipelines, reference genome versions, and quality control practices, all of which can hinder data integration in agri-genomics. Although individual projects have unique goals, implementing uniform and harmonized documentation enables other data users to replicate procedures with both new and historical genomic data, supporting effective data sharing and reuse beyond original project goals. This recommendation is aligned with the BioCompute Object (formalized as IEEE 2791), which focuses on improving communication, provenance, and reproducibility of bioinformatics workflows [79], [80].

Aside from commercial laboratories that deliver specific services and have already implemented standardized quality control for generating genomic data, this recommendation is expected to benefit all

the other relevant parties (academia, industry, and industry–academia groups) working with genomic data, as the objective is to establish standardized protocols and documentation to ensure data practices are well annotated and consistent among parties.

Recommendation 4: Create a Shared Discovery Catalogue and Repository Policy

It is recommended to develop a repository-plus-catalogue model, avoiding the attempt to centralize all agri-genomic data in a single storage location. Data could be deposited in international and well-established public repositories at the same time as they are registered to a Canadian discovery catalogue. This approach promotes the use of a unified index, with persistent identifiers, machine-readable metadata, links to publications, and controlled access. Together, this will support data discoverability while supporting long-term data preservation, data ownership, and sovereignty.

This recommendation is expected to benefit academia, industry, and industry–academia parties.

Recommendation 5: Enable Cross-Project Data Integration and Reuse

It is recommended to establish minimum integration specifications and to adopt consistent naming conventions, shared genome reference versions, tools for mapping identifiers, quality control validation utilities, APIs for data discovery, and practical integration guidelines. In addition, researchers are encouraged to contribute by providing example pipelines, demonstrating how their datasets can be integrated with those generated by other user groups. A standard cross-project integration approach such as this will provide essential support for data interoperability. By implementing these practices, data sharing and reuse will be facilitated for all parties involved in genomic data generation.

This recommendation is expected to benefit academia, industry, and industry–academia groups by providing standardized support that enhances data interoperability among these sectors.

Recommendation 6: Support Training and Capacity Building

It is recommended that curricula include data standards principles and be created and made available by a committee, with the goal of providing ongoing or on-demand training modules. Prioritizing investments in the development of highly qualified personnel with expertise in genomic data is essential for the sustained adoption and maintenance of standardized data practices. Implementing a train-the-trainer model could help further build internal capacity. Establishing a comprehensive training ecosystem would connect various groups and institutions, reduce technical barriers, and enhance data quality.

Promoting training and capacity building is expected to have a positive impact on all relevant parties (academia, industry, and industry–academia groups) involved with genomic data, resulting in a greater number of qualified personnel able to collaborate across parties, thereby supporting the implementation of best practices.

Recommendation 7: Implement Monitoring, Evaluation, and Refinement

It is recommended to implement ongoing monitoring, evaluation, and refinement of data processes and practices, and to establish clear metrics for adoption, quality, reuse, timeliness, and catalogue completeness to support effective implementation and lasting adoption of standardized practices. This approach could be used to assess progress and identify barriers, while ensuring that updates in practices and changes in processes remain transparent, with reduced disruption to workflows.

This recommendation is expected to benefit all relevant parties (academia, industry, commercial laboratories, and industry–academia groups) working with genomic data, resulting in more harmonized data practices across all sectors while improving data documentation, which will in turn improve usability and support long-term integration of agri-genomic data.

Recommendation 8: Secure Funding for Long-Term Data Preservation

It is recommended to allocate dedicated budgets or funding streams both within and outside projects for data archiving, metadata curation, repository fees, data quality control/integrity, and legacy backfilling. Sustained funding for long-term data preservation is essential for enabling data sharing and reuse once projects are completed, while minimizing the risk of data loss or inaccessibility. As part of a data management plan, funding agencies could require information related to long-term preservation, including details about short- and long-term responsibilities, data access dependencies, retention timeline, and internal verification procedures.

This recommendation is expected to benefit academia, industry, and industry–academia groups. Dedicated funding will support consistent archiving, harmonizing, and curation, enabling these parties to reuse and integrate data across projects and timeframes. Further benefits include improved data integrity and reduced likelihood of data loss, ensuring that valuable agri-genomic datasets remain available for future scientific discovery and commercial applications.

In conclusion, effective standardization must address every stage of the data lifecycle, with metadata harmonization anchored in scientific community standards and serving as the foundation for all other standardization efforts and requirements. This process cannot rely solely on technical or general tools, as it requires coherent governance and management structures established collaboratively by a group of experts working toward a common goal. This group could also play a critical role in clarifying governance expectations on sensitive topics, such as data ownership, sovereignty, data access conditions, data sharing agreement terms, and both short- and long-term stewardship responsibilities. Doing so will help ensure data integrity and accessibility are maintained even after projects conclude. For long-term preservation and reusability, a hybrid model could be applied for data storage, using international repositories while maintaining a Canadian discovery catalogue. This standardization process would be further facilitated with investments in training and capacity building, supported by dedicated long-term funding for data preservation.

Based on this report's findings from the literature review and landscape analysis, national and international discussions, and observed practices from the ICT screening, it is evident that Canada is well-positioned to implement agri-genomic data standards at the national level. Moreover, there is potential to extend the application of these standards beyond genomics climate mitigation efforts to include all types of genomic data generated across Canada. While some factors may limit or slow the widespread adoption of standards, the foundational elements are already established, with strong scientific capacity, collaborative research networks (represented by the ICTs), and alignment of ICTs with international practices. To begin to address standards in agri-genomics data, the adoption of unified metadata templates, coordinated governance, robust training, and long-term stewardship should be prioritized. Through these efforts, Canada could develop a genomic data ecosystem that is interoperable and reusable, and that recognizes Canadian researchers' contributions to genomic dataset generation and their role in advancing climate-smart agriculture innovations, both nationally and internationally.

References

[1] J. I. Marsh, H. Hu, M. Gill, J. Batley, and D. Edwards, "Crop breeding for a changing climate: integrating phenomics and genomics with bioinformatics," *Theor. Appl. Genet.,* vol. 134, no. 6, pp. 1677–1690, 2021, https://doi.org/10.1007/s00122-021-03820-3.

[2] D. P. Berry, "Genomic information in livestock has multiple uses in precision breeding and management," *Livestock,* vol. 24, no. 1, pp. 30–33, 2019, https://doi.org/10.12968/live.2019.24.1.30.

[3] National Human Genome Research Institute. "Genetics vs. genomics fact sheet." Accessed: Mar. 29, 2026. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

[4] M. W. Bevan, C. Uauy, B. B. H. Wulff, J. Zhou, K. Krasileva, and M. D. Clark, "Genomic innovation for crop improvement," *Nature,* vol. 543, no. 7645, pp. 346–354, 2017, https://doi.org/10.1038/nature22011.

[5] M. Georges, C. Charlier, and B. Hayes, "Harnessing genomic information for livestock improvement," *Nat. Rev. Genet.,* vol. 20, no. 3, pp. 135–156, 2019, https://doi.org/10.1038/s41576-018-0082-2.

[6] J. Crossa *et al.,* "Genomic selection in plant breeding: methods, models, and perspectives," *Trends Plant Sci.,* vol. 22, no. 11, pp. 961–975, 2017, https://doi.org/10.1016/j.tplants.2017.08.011.

[7] Y. Xu *et al.,* "Enhancing genetic gain through genomic selection: from livestock to plants," *Plant Commun.,* vol. 1, no. 1, 2020, Art. no. 100005, https://doi.org/10.1016/j.xplc.2019.100005.

[8] P. Trivedi, J. E. Leach, S. G. Tringe, T. Sa, and B. K. Singh, "Plant–microbiome interactions: from community assembly to plant health," *Nat. Rev. Microbiol.,* vol. 18, no. 11, pp. 607–621, 2020, https://doi.org/10.1038/s41579-020-0412-1.

[9] C. D. Gutierrez Reyes *et al.,* "Multi omics applications in biological systems," *Curr. Issues Mol. Biol.,* vol. 46, no. 6, pp. 5777–5793, 2024, https://doi.org/10.3390/cimb46060345.

[10] E. Clagnan *et al.,* "Culturomics- and metagenomics-based insights into the soil microbiome preservation and application for sustainable agriculture," *Front. Microbiol.,* vol. 15, 2024, Art. no. 1473666, https://doi.org/10.3389/fmicb.2024.1473666.

[11] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS Comput. Biol.,* vol. 13, no. 5, 2017, Art. no. e1005457, https://doi.org/10.1371/journal.pcbi.1005457.

[12] R. K. Varshney, A. Bohra, J. Yu, A. Graner, Q. Zhang, and M. E. Sorrells, "Designing future crops: genomics-assisted breeding comes of age," *Trends Plant Sci.,* vol. 26, no. 6, pp. 631–649, 2021, https://doi.org/10.1016/j.tplants.2021.03.010.

[13] M. A. Hossain and H. A. Roslan, "Editorial: Emerging genomic technologies for agricultural biotechnology: current trends and future prospects," *Front. Plant Sci.,* vol. 14, 2023, Art. no. 1263289, https://doi.org/10.3389/fpls.2023.1263289.

[14] D. M. Leigh *et al.,* "Best practices for genetic and genomic data archiving," *Nat. Ecol. Evol.,* vol. 8, no. 7, pp. 1224–1232, 2024, https://doi.org/10.1038/s41559-024-02423-7.

[15] Genome Canada, "Climate-smart agriculture and food systems: Interdisciplinary Challenge Teams funding opportunity," Genome Canada, Ottawa, ON, Canada, Dec. 2022. Accessed: Mar. 29, 2026. [Online]. Available: https://genomecanada.ca/wp-content/uploads/2022/12/Climate-Smart-Agriculture-and-Food-Systems_ICT-Funding-Opportunity-3.pdf

[16] R. Yotova and B. M. Knoppers, "The right to benefit from science and its implications for genomic data sharing," *Eur. J. Int. Law,* vol. 31, no. 2, pp. 665–691, 2020, https://doi.org/10.1093/ejil/chaa028.

[17] C. K. Tuggle *et al.,* "Current challenges and future of agricultural genomes to phenomes in the USA," *Genome Biol.,* vol. 25, 2024, Art. no. 8, https://doi.org/10.1186/s13059-023-03155-w.

[18] M. D. Wilkinson *et al.,* "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data,* vol. 3, 2016, Art. no. 160018, https://doi.org/10.1038/sdata.2016.18.

[19] Agriculture and Agri-Food Canada. "Agriculture and Agri-Food Canada's Open Science Action Plan 2021–2022 to 2025–2026 – Enable and equip." Accessed: Mar. 29, 2026. [Online]. Available: https://agriculture.canada.ca/en/science/open-data-and-scientific-agricultural-publications/agriculture-and-agri-food-canadas-open-science-action-plan-2021-2022-2025-2026-enable-and-equip

[20] Science.gc.ca. "Tri-Agency Research Data Management Policy." Accessed: Mar. 29, 2026. [Online]. Available: https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy

[21] Genome Canada, "Data release and resource sharing policy," Genome Canada, Ottawa, ON, Canada, 2023. Accessed: Mar. 29, 2026. [Online]. Available: https://genomecanada.ca/wp-content/uploads/2022/05/gcdatasharingpolicies16-09-23-1.pdf

[22] National Agricultural Library. "Ag Data Commons collection policies." Accessed: Mar. 30, 2026. [Online]. Available: https://www.nal.usda.gov/services/agdatacommons/policies

[23] Environment and Natural Resourcess Canada. "Net-zero emissions by 2050." Accessed: Mar. 30, 2026. [Online]. Available: https://www.canada.ca/en/services/environment/weather/climatechange/climate-plan/net-zero-emissions-2050.html

[24] E. W. Sayers *et al.*, "Database resources of the National Center for Biotechnology Information in 2023," *Nucleic Acids Res.,* vol. 51, no. D1, pp. D29–D38, 2023, https://doi.org/10.1093/nar/gkac1032.

[25] L. Harper *et al.,* "AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture," Database, vol. 2018, 2018, Art. no. bay088, https://doi.org/10.1093/database/bay088.

[26] International Society for Biocuration, "Biocuration: Distilling data into knowledge," *PLoS Biol.,* vol. 16, no. 4, 2018, Art. no. e2002846, https://doi.org/10.1371/journal.pbio.2002846.

[27] J. Chamorro-Padial, J. Virgili-Gomá, R. Gil, M. Teixidó, and R. García, "Agriculture data sharing review," *Heliyon,* vol. 11, no. 1, 2025, Art. no. e41109, https://doi.org/10.1016/j.heliyon.2024.e41109.

[28] J. L. Clarke *et al.,* "Data sharing and ontology use among agricultural genetics, genomics, and breeding databases and resources of the Agbiodata consortium," Database, vol. 2023, 2023, Art. no. baad076, https://doi.org/10.1093/database/baad076.

[29] K. E. Lotterhos, "Principles in experimental design for evaluating genomic forecasts," *Methods Ecol. Evol.,* vol. 15, no. 9, pp. 1466–1482, 2024, https://doi.org/10.1111/2041-210X.14379.

[30] S. Mukherjee *et al.,* "Genomes OnLine Database (GOLD) v.10: New features and updates," *Nucleic Acids Res.,* vol. 53, no. D1, pp. D989–D997, 2025, https://doi.org/10.1093/nar/gkae1000.

[31] E. M. Wood-Charlson *et al.,* "The National Microbiome Data Collaborative: Enabling microbiome science," *Nat. Rev. Microbiol.,* vol. 18, no. 6, pp. 313–314, 2020, https://doi.org/10.1038/s41579-020-0377-0.

[32] L. Thompson *et al.,* "Introduction to metadata and ontologies: Everything you always wanted to know about metadata and ontologies (but were afraid to ask)," National Microbiome Data Collaborative, Berkeley, CA, United States, Mar. 2020. [Online]. Available: https://doi.org/10.25979/1607365.

[33] B. Sibbett, L. H. Rieseberg, and S. Narum, "The Genomic Observatories Metadatabase," *Mol. Ecol. Resour.,* vol. 20, no. 6, pp. 1453–1454, 2020, https://doi.org/10.1111/1755-0998.13283.

[34] D. Zurell *et al.,* "A standard protocol for reporting species distribution models," *Ecography,* vol. 43, no. 9, pp. 1261–1277, 2020, https://doi.org/10.1111/ecog.04960.

[35] H. A. Lewin *et al.,* "Earth BioGenome Project: Sequencing life for the future of life," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 115, no. 17, pp. 4325–4333, 2018, https://doi.org/10.1073/pnas.1720115115.

[36] M. K. N. Lawniczak *et al.,* "Standards recommendations for the Earth BioGenome Project," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 119, no. 4, 2022, Art. no. e2115639118, https://doi.org/10.1073/pnas.2115639118.

[37] D. Poulin-Laprade et al., "Requirements and considerations for effective implementation of integrated One Health antimicrobial resistance research," *Can. J. Microbiol.,* vol. 71, pp. 1–17, 2025, https://doi.org/10.1139/cjm-2024-0194.

[38] E. J. Griffiths *et al.,* "Crossing the streams: Improving data quality and integration across the One Health genomics continuum with data standards and implementation strategies," *Can. J. Microbiol.,* vol. 71, pp. 1–14, 2025, https://doi.org/10.1139/cjm-2024-0203.

[39] A. Kotal, L. Elluri, D. Gupta, V. Mandalapu, and A. Joshi, "Privacy-preserving data sharing in agriculture: Enforcing policy rules for secure and confidential data synthesis," in *Proc. 2023 IEEE Int. Conf. Big Data (BigData),* Sorrento, Italy, Dec. 15–18, 2023, pp. 5519–5528. https://doi.org/10.1109/BigData59044.2023.10386276.

[40] European Parliament and Council of the European Union, 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Accessed: Mar. 30, 2026. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[41] A. A. Taddese, A. C. Addis, and B. T. Tam, "Data stewardship and curation practices in AI-based genomics and automated microscopy image analysis for high-throughput screening studies: Promoting robust and ethical AI applications," *Hum. Genomics,* vol. 19, no. 1, 2025, Art. no. 16, https://doi.org/10.1186/s40246-025-00716-x.

[42] J. C. Wilgenbusch, P. G. Pardey, N. Hospodarsky, and B. J. Lynch, "Addressing new data privacy realities affecting agricultural research and development: A tiered-risk, standards-based approach," *Agron. J.,* vol. 114, no. 5, pp. 2653–2668, 2022, https://doi.org/10.1002/agj2.20968.

[43] I. Keenum, S. A. Jackson, E. Eloe-Fadrosh, and L. M. Schriml, "A standards perspective on genomic data reusability and reproducibility," *Front. Bioinform.,* vol. 5, 2025, Art. no. 1572937, https://doi.org/10.3389/fbinf.2025.1572937.

[44] M. Sheikh, F. Iqra, H. Ambreen, K. A. Pravin, M. Ikra, and Y. S. Chung, "Integrating artificial intelligence and high-throughput phenotyping for crop improvement," *J. Integr. Agric.,* vol. 23, no. 6, pp. 1787–1802, 2024, https://doi.org/10.1016/j.jia.2023.10.019.

[45] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, and M. van der Velde, "Monitoring crop phenology with street-level imagery using computer vision," *Comput. Electron. Agric.,* vol. 196, 2022, Art. no. 106866, https://doi.org/10.1016/j.compag.2022.106866.

[46] International Committee for Animal Recording. "The global standard for livestock data." Accessed: Mar. 30, 2026. [Online]. Available: https://www.icar.org

[47] ICAR Wiki. "ICAR Guidelines." Accessed: Mar. 30, 2026. [Online]. Available: https://wiki.icar.org/index.php/Guidelines

[48] M. Kapoor *et al.,* "Building a FAIR data ecosystem for incorporating single-cell transcriptomics data into agricultural genome to phenome research," *Front. Genet.,* vol. 15, 2024, Art. no. 1460351, https://doi.org/10.3389/fgene.2024.1460351.

[49] N. L. Anglin, P. Wenzl, V. Azevedo, C. Lusty, D. Ellis, and D. Gao, "Genotyping genebank collections: Strategic approaches and considerations for optimal collection management," *Plants,* vol. 14, no. 2, 2025, Art. no. 252, https://doi.org/10.3390/plants14020252.

[50] T. G. Brooks, N. F. Lahens, A. Mrčela, and G. R. Grant, "Challenges and best practices in omics benchmarking," *Nat. Rev. Genet.,* vol. 25, no. 5, pp. 326–339, May 2024, https://doi.org/10.1038/s41576-023-00679-6.

[51] A. Marrano *et al.,* "A teaching and training framework to promote findable, accessible, interoperable, and reusable data generation in agriculture," *Database,* vol. 2025, 2025, Art. no. baaf034, https://doi.org/10.1093/database/baaf034.

[52] E. K. Moore *et al.,* "Agricultural data management and sharing: Best practices and case study," *Agron. J.,* vol. 114, no. 5, pp. 2624–2634, 2022, https://doi.org/10.1002/agj2.20639.

[53] A. Hafner *et al.,* "Data reuse in agricultural genomics research: Challenges and recommendations," *GigaScience,* vol. 14, 2025, Art. no. giae106, https://doi.org/10.1093/gigascience/giae106.

[54] C. Tenopir *et al.,* "Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide," *PLoS One,* vol. 15, no. 3, 2020, Art. no. e0229003, https://doi.org/10.1371/journal.pone.0229003.

[55] D. Field *et al.,* "The minimum information about a genome sequence (MIGS) specification," *Nat. Biotechnol.,* vol. 26, no. 5, pp. 541–547, 2008, https://doi.org/10.1038/nbt1360.

[56] A. Ault, S. Palacios, and J. Evans, "Agriculture data sharing: Conceptual tools in the technical toolbox and implementation in the Open Ag Data Alliance framework," *Agron. J.,* vol. 114, no. 5, pp. 2681–2692, 2022, https://doi.org/10.1002/agj2.21007.

[57] The FAANG Consortium *et al.,* "Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project," *Genome Biol., v*ol. 16, no. 1, 2015, Art. no. 57, https://doi.org/10.1186/s13059-015-0622-4.

[58] J. B. Byrd, A. C. Greene, D. V. Prasad, X. Jiang, and C. S. Greene, "Responsible, practical genomic data sharing that accelerates research," *Nat. Rev. Genet.,* vol. 21, no. 10, pp. 615–629, 2020, https://doi.org/10.1038/s41576-020-0257-5.

[59] D. G. E. Gomes *et al.,* "Why don't we share data and code? Perceived barriers and benefits to public archiving practices," *Proc. R. Soc. B: Biol. Sci.,* vol. 289, no. 1987, 2022, Art. no. 20221113, https://doi.org/10.1098/rspb.2022.1113.

[60] O. Zafar, R. S. Gonzalez, G. Wilkins, A. Morales, and E. Ayday, "Privacy-preserving data linkage across private and public datasets for collaborative agriculture research," Sep. 9, 2024, *arXiv:2409.06069.* Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2409.06069

[61] N. Waithira, B. Mutinda, and P. Y. Cheah, "Data management and sharing policy: The first step towards promoting data sharing," *BMC Med.,* vol. 17, no. 1, 2019, Art. no. 80, https://doi.org/10.1186/s12916-019-1315-8.

[62] A. Bernasconi, "Data quality-aware genomic data integration," *Comput. Methods Programs Biomed. Update,* vol. 1, 2021, Art. no. 100009, https://doi.org/10.1016/j.cmpbup.2021.100009.

[63] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics,* vol. 157, no. 4, pp. 1819–1829, 2001, https://doi.org/10.1093/genetics/157.4.1819.

[64] M. De Donato, S. O. Peters, S. E. Mitchell, T. Hussain, and I. G. Imumorin, "Genotyping-by-sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing," *PLoS One,* vol. 8, no. 5, 2013, Art. no. e0062137, https://doi.org/10.1371/journal.pone.0062137.

[65] J. Lee, J. H. Jung, and S.-H. Oh, "Enhancing animal breeding through quality control in genomic data – A review," *J. Anim. Sci. Technol.,* vol. 66, no. 6, pp. 1099–1108, 2024, https://doi.org/10.5187/jast.2024.e92.

[66] W. Hemstrom, J. A. Grummer, G. Luikart, and M. R. Christie, "Next-generation data filtering in the genomics era," *Nat. Rev. Genet.,* vol. 25, no. 11, pp. 750–767, 2024, https://doi.org/10.1038/s41576-024-00738-6.

[67] S.-A. Sansone *et al.,* "FAIRsharing as a community approach to standards, repositories and policies," *Nat. Biotechnol.,* vol. 37, no. 4, pp. 358–367, 2019, https://doi.org/10.1038/s41587-019-0080-8.

[68] C. Hedeler and P. Missier, "Information quality management challenges for high-throughput data," in *Biological Database Modeling,* J. Chen and A. S. Sidhu, Eds., Norwood, MA, United States: Artech House, 2008, ch. 6, pp. 81–98. [Online]. Available: https://books-library.website/files/books-library.online-02092324Bp7Z0.pdf

[69] F. Moussouni and L. Berti-Équille, "Cleaning, Integrating, and warehousing genomic data from biomedical resources," in *Biological Knowledge Discovery Handbook,* M. Elloumi and A. Y. Zomaya, Eds., Hoboken, NJ, United States: Wiley, 2013, ch. 2, pp. 35–58. [Online]. Available: https://doi.org/10.1002/9781118617151.ch02

[70] D. Field *et al.,* "The Genomic Standards Consortium," *PLoS Biol.,* vol. 9, no. 6, 2011, Art. no. e1001088, https://doi.org/10.1371/journal.pbio.1001088.

[71] CNCB-NGDC Members and Partners, "Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2022," *Nucleic Acids Res.,* vol. 50, no. D1, pp. D27–D38, 2022, https://doi.org/10.1093/nar/gkab951.

[72] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nat. Protoc.,* vol. 5, no. 9, pp. 1564–1573, 2010, https://doi.org/10.1038/nprot.2010.116.

[73] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet J.,* vol. 17, no. 1, pp. 10–12, 2011, https://doi.org/10.14806/ej.17.1.200.

[74] C. Camacho *et al.,* "BLAST+: Architecture and applications," *BMC Bioinform.,* vol. 10, no. 1, 2009, Art. no. 421, https://doi.org/10.1186/1471-2105-10-421.

[75] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods,* vol. 9, no. 4, pp. 357–359, 2012, https://doi.org/10.1038/nmeth.1923.

[76] M. O. Adetunji, S. J. Lamont, B. Abasht, and C. J. Schmidt, "Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data," *PLoS One,* vol. 14, no. 9, 2019, Art. no. e0216838, https://doi.org/10.1371/journal.pone.0216838.

[77] S. Pavan, C. Delvento, L. Ricciardi, C. Lotti, E. Ciani, and N. D'Agostino, "Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies," *Front. Genet.,* vol. 11, 2020, Art. no. 447, https://doi.org/10.3389/fgene.2020.00447.

[78] M. Baker, "1,500 scientists lift the lid on reproducibility," Nature, vol. 533, no. 7604, pp. 452–454, 2016, https://doi.org/10.1038/533452a.

[79] V. Simonyan, J. Goecks, and R. Mazumder, "Biocompute objects—A step towards evaluation and validation of biomedical scientific computations," *PDA J. Pharm. Sci. Technol.,* vol. 71, no. 2, pp. 136–146, 2017, https://doi.org/10.5731/pdajpst.2016.006734.

[80] *IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication,* IEEE 2791, Institute of Electrical and Electronics Engineers, Piscataway, NJ, United States, 2020. Accessed: Apr. 8, 2026. [Online]. Available: https://standards.ieee.org/ieee/2791/7337

[81] K. C. O'Doherty *et al.,* "Toward better governance of human genomic data," *Nat. Genet.,* vol. 53, no. 1, pp. 2–8, 2021, https://doi.org/10.1038/s41588-020-00742-6.

[82] S. R. Carroll *et al.,* "The CARE Principles for Indigenous Data Governance," *Data Sci. J.,* vol. 19, no. 1, 2020, Art. no. 43, https://doi.org/10.5334/dsj-2020-043.

[83] P. Krajewski *et al.,* "Towards recommendations for metadata and data handling in plant phenotyping," *J. Exp. Bot.,* vol. 66, no. 18, pp. 5417–5427, 2015, https://doi.org/10.1093/jxb/erv271.

[84] M. Buck and C. Hamilton, "The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity," *Rev. Eur. Community Int. Environ. Law,* vol. 20, no. 1, pp. 47–61, 2011, https://doi.org/10.1111/j.1467-9388.2011.00703.x.

[85] S. C. Dyer *et al.,* "Ensembl 2025," *Nucleic Acids Res.,* vol. 53, no. D1, pp. D948–D957, 2025, https://doi.org/10.1093/nar/gkae1071.

[86] D. M. Bolser, D. M. Staines, E. Perry, and P. J. Kersey, "Ensembl Plants: Integrating tools for visualizing, mining, and analyzing plant genomic data," in *Plant Genomics Databases: Methods and Protocols,* vol. 1533, A. D. J. van Dijk, Ed., New York, NY, United States: Humana Press, 2017, pp. 1–31. [Online]. Available: https://doi.org/10.1007/978-1-4939-6658-5_1

[87] National Animal Genome Research Program. "Animal Quantitative Trait Loci Database (Animal (QTLdb)." Accessed: Mar. 31, 2026. [Online]. Available: https://www.animalgenome.org/cgi-bin/QTLdb/index

[88] Z. L. Hu, C. A. Park, and J. M. Reecy, "Bringing the Animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services," *Nucleic Acids Res.,* vol. 50, no. D1, pp. D956–D961, 2022, https://doi.org/10.1093/nar/gkab1116.

[89] Global Agricultural Research Data Innovation and Acceleration Network. "Welcome to GARDIAN: Discover and access agri-food research assets." Accessed: Mar. 30, 2026. [Online]. Available: https://gardian.cgiar.org

[90] S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, A. R. Martin, and K. Kuchenbaecker, "A roadmap to increase diversity in genomic studies," *Nat. Med.,* vol. 28, no. 2, pp. 243–250, 2022, https://doi.org/10.1038/s41591-021-01672-4.

[91] "100,000 genomes — in Africa, for Africa," *Nat. Biotechnol.,* vol. 40, no. 7, p. 988, 2022, https://doi.org/10.1038/s41587-022-01404-0.

[92] VistaMilk. "VistaMilk SFI Research Centre." Accessed: Mar. 30, 2026. [Online]. Available: https://www.vistamilk.ie

[93] University of Tennessee Beef Genomics. "The Rowan Lab." Accessed: Mar. 30, 2026. [Online]. Available: https://sites.google.com/utk.edu/rowanlab/home

[94] Helical. "Helical: The genetics operating system." Accessed: Mar. 30, 2026. [Online]. Available: https://www.helicalco.com

[95] Environment and Climate Change Canada, "National inventory report 1990–2019: Greenhouse gas sources and sinks in Canada, Part 1," Environment and Climate Change Canada, Gatineau, QC, Canada, 2021. Accessed: Mar. 30, 2026. [Online]. Available: https://publications.gc.ca/collections/collection_2021/eccc/En81-4-2019-1-eng.pdf

[96] Environment and Climate Change Canada, "National inventory report 1990–2020: Greenhouse gas sources and sinks in Canada," Environment and Climate Change Canada, Gatineau, QC, Canada, 2022. Accessed: Mar. 29, 2026. [Online]. Available: https://publications.gc.ca/collections/collection_2022/eccc/En81-4-2020-1-eng.pdf

[97] Statistics Canada. "Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units." Accessed: Mar. 30, 2026. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210035901

[98] Agriculture and Agri-Food Canada. "What we heard report: Fertilizer emissions reduction." Accessed: Mar. 29, 2026. [Online]. Available: https://agriculture.canada.ca/en/department/transparency/public-opinion-research-consultations/share-ideas-fertilizer-emissions-reduction-target/what-we-heard-report-fertilizer-emissions-reduction

[99] I. Shcherbak, N. Millar, and G. P. Robertson, "Global metaanalysis of the nonlinear response of soil nitrous oxide (N2O) emissions to fertilizer nitrogen," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 111, no. 25, pp. 9199–9204, 2014, https://doi.org/10.1073/pnas.1322434111.

[100] Genome Alberta. "Leveraging genomics to achieve dairy net zero." [Online]. Available: https://genomealberta.ca/project/leveraging-genomics-to-achieve-dairy-net-zero/

[101] Standards Council of Canada. "Sub-programs: Verified Carbon Standard." Accessed: Mar. 30, 2026. [Online]. Available: https://scc-ccn.ca/accreditation-scheme/verified-carbon-standard

[102] CarbonClear Canada. "CarbonClear Canada: Offset your emissions with Canadian carbon certified credits." Accessed: Mar. 30, 2026. [Online]. Available: https://carbonclear.ca

[103] Bovine Long Read Consortium Quality Control Subcommittee. "Bovine Long Read Consortium guidelines for run 1." Accessed: Mar. 30, 2026. [Online]. Available: https://github.com/BovLRC/bovLRC-documentation/blob/main/document.md

[104] Ontario Genomics, "Food innovation in Canada report: An $18.8 billion opportunity in precision fermentation and cultivated foods," Ontario Genomics, Toronto, ON, Canada, 2025. Accessed: Mar. 30, 2026. [Online]. Available: https://www.ontariogenomics.ca/food-innovation-in-canada-report-2025

[105] R. Eibl, Y. Senn, G. Gubser, V. Jossen, C. van den Bos, and D. Eibl, "Cellular agriculture: opportunities and challenges," *Annu. Rev. Food Sci. Technol.,* vol. 12, no. 1, pp. 51–73, 2021, https://doi.org/10.1146/annurev-food-063020-123940.

[106] M. Goswami *et al.,* "Cell-cultivated aquatic food products: Emerging production systems for seafood," *J. Biol. Eng.,* vol. 18, no. 1, 2024, Art. no. 43, https://doi.org/10.1186/s13036-024-00436-1.

[107] G. S. Barsh *et al.,* "PLoS Genetics data sharing policy: In pursuit of functional utility," *PLoS Genet.,* vol. 11, no. 12, 2015, Art. no. e1005716, https://doi.org/10.1371/journal.pgen.1005716.

[108] Springer Nature. "Research data policy: Data availability statements." Accessed: Mar. 30, 2026. [Online]. Available: https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements

[109] Ag Data Commons. "Browse – Ag Data Commons research." Accessed: Mar. 30, 2026. [Online]. Available: https://agdatacommons.nal.usda.gov

[110] National Agricultural Library. "Ag Data Commons user guide: Data submission manual." Accessed: Mar. 31, 2026. [Online]. Available: https://www.nal.usda.gov/services/agdatacommons/data-submission-manual

[111] World Wide Web Consortium. "Data Catalog Vocabulary (DCAT) – Version 3." Accessed: Mar. 30, 2026. [Online]. Available: https://www.w3.org/TR/vocab-dcat-3

[112] Resources.data.gov. "DCAT-US Schema v1.1 (Project Open Data Metadata Schema)." Accessed: Mar. 30, 2026. [Online]. Available: https://resources.data.gov/resources/dcat-us

[113] National Agricultural Library. "National Agricultural Library." Accessed: Mar. 30, 2026. [Online]. Available: https://www.nal.usda.gov

[114] J. Ingwersen. "USDA migrates data archive to new website, dropping Cornell's Mann Library." Reuters. Accessed: Mar. 31, 2026. [Online]. Available: https://www.reuters.com/world/us/usda-migrates-data-archive-new-website-dropping-cornells-mann-library-2025-09-30

[115] United States Department of Agriculture, "Fiscal year 2024–2026 data strategy," United States Department of Agriculture, Washington, DC, United States, 2023. Accessed: Mar. 31, 2026. [Online]. Available: https://www.usda.gov/sites/default/files/documents/fy-2024-2026-usda-data-strategy.pdf

Appendix A—Questionnaire for the International Interviews

The following questions were developed to identify the best processes and practices for data collection, management, analysis, reporting, sharing, and governance in agricultural genomics.

Data, Infrastructure and Technologies

1. What infrastructures are available for processes related to genomic data generation?
2. What are your practices for genomic data collection?
3. What are the main challenges, roadblocks, hurdles or gaps in generating genomic data (affecting aspects such as consistency, accuracy, and/or quality)? Are there any tools and/or strategic plans to help address the challenges?
4. What genomic data types have been generated over the years of the project? If possible, specify the platforms used.
5. What are the intended uses of each genomic data type?

Third-Party Labs and Data Transfer

1. Have you used a third-party lab to generate genomic data? If so, how are the genomic data files transferred to the research team?
2. What are the main challenges related to genomic data transfer from lab and/or service providers to the research team?

Data Management and Sharing Practices

1. What are your practices for data management, storage and sharing? Is there any standard operating procedure in place?
2. Which international platforms have been used for genomic data storage and sharing (e.g., NCBI)?
3. Has funding been provided to properly set up archiving project data for future use?
4. Is there a designated person to oversee raw files, implement standard initial quality control, and handle data throughout the study period?
5. How often is the genomic dataset archive updated (e.g., monthly, quarterly)?
6. As several researchers and groups are enrolled in the project, do you have an internal policy for data sharing across groups?
7. Have you identified any problems affecting genomic data comparability across different projects? How were these issues addressed?
8. Do you have a long-term plan for data management and storage, especially considering data management, governance, and sharing when project funds are terminated?

Training and Skill Development

1. Is there any training for data generators in data management? How are genomic data management, analysis, and sharing skills improved throughout the project?
2. Do you have a training program for new team members, undergraduate, and/or graduate students related to data management, data quality control, and data analysis?
3. Is there any training on standardized sample identifiers and maintaining crucial data associated with metadata?

Data Review and Metadata Preparation

1. Is there a designated person to review the dataset once it is submitted (for storage or manuscript publication) to ensure files are not corrupt, contain the appropriate information, and basic metadata?
2. Do you have any standard operating procedure or template for metadata preparation?
3. Is there any standard operating procedure, template, or standard format for data identification in place?

Reporting and Communication

1. What are your practices in reporting genomics results (e.g., manuscripts, white papers, technical notes, interviews, social media, project website, etc.)?
2. What have you identified as the most effective reporting strategy?

Data Reuse and FAIR Principles

1. Can your data be integrated (internally and internationally) and reused for discovering new knowledge? If so, what would be the potential challenges in combining and comparing with other datasets?
2. How likely is your genomic data to be reused internally and internationally?
3. What are your strategies to leverage data reuse internally and internationally? Is there any strategy in place?
4. Are the FAIR principles (findability, accessibility, interoperability, and reusability) implemented? If so, how do you ensure that?
5. Is genomic data previously collected by other projects being incorporated? If so, have there been issues utilizing that data or are there processes in place for integration?

Support and Documentation

1. Is there any infrastructure to support data sharing, data quality checks, data reformatting, curation, and re-analysis?
2. Are there any documentation or processes that will help with clearly connecting genomic data and phenotypes?

Appendix B—Questionnaire for the Interdisciplinary Challenge Teams

1. **Data, Infrastructure and Technologies**

    a. Of the following list, what types of genomic data will be generated from your project: array genotypes, (whole) genome sequence, (whole) exome sequence, RNA seq., 16s/18s metagenome, (shotgun) whole metagenome, and/or other (please specify)?

        i. Please specify the platforms used.

        ii. What are the intended uses of each genomic data type?

        iii. Describe the challenges/roadblocks/gaps in using the genomic data.

    b. What infrastructures are available to your group to support genomic data generation, including outsourcing?

    c. What are the main challenges, roadblocks, hurdles or gaps in generating genomic data (e.g., consistency, accuracy, quality)?

        i. Are there any tools or strategic plan in place to address these challenges?

        ii. What is the expected genomic data size upon completion of the project, if known?

2. **Third-Party Labs and Data Transfer**

    a. If a third-party lab was contracted to generate genomic data, how are the resulting data files transferred to the research team?

    b. What are the main challenges related to transferring genomic data from the third-party lab or service providers to the research team?

    c. If multiple labs were contracted to generate genomic data, describe any challenges related to data comparability.

3. **Data Management and Sharing Practices**

    a. What are your general practices for data management, storage and sharing?

        i. Is there a standard operating procedure and/or a dedicated team member to oversee?

        ii. Are there any ISO standards in place for data management and sharing? Please specify.

        iii. How often is the project dataset archive updated (e.g., monthly, quarterly)?

    b. Will any international platforms be used for genomic data storage and sharing (e.g., NCBI)?

    c. Given the involvement of multiple researchers and groups, is there an internal policy for data sharing across teams?

d. Have you encountered any issues affecting genomic data comparability across different groups and/or projects?
    i. If so, how were they addressed?
    ii. Are there standards or processes in place for data integration?

e. Do you have a long-term plan for data management, storage and governance, especially after project funding ends?
    i. Has funding been provided to properly set up data archiving for future use?
    ii. What challenges could affect future data re-use?

4. Data Review and Metadata Preparation

a. Do you have a standard operating procedure or template for genomic data identification and for metadata preparation?

b. Specify the metadata for/to the genomic data in your project.

c. In addition to storage, is there a designated person responsible for reviewing datasets upon submission to ensure files are not corrupt, contain the appropriate information, and the metadata?

5. Reporting and Communication

a. What are your practices in reporting genomics results (e.g., raw data and processed data deposits, manuscripts, white papers, technical notes, interviews, social media, project website, etc.)?
    i. What reporting strategy has proven most effective?

b. Have any policies or regulations affected any form of your project reporting?

c. Is your genomic data relevant to any of the following, and how so (if applicable)?
    i. Regulators
    ii. Policy-makers
    iii. Industry
    iv. General consumers

d. For the purpose of decision making by these four groups, what is the appropriate presentation of genomic data, if any?
    i. Regulators
    ii. Policy-makers
    iii. Industry
    iv. General consumers

e. Is there influence of a policy and/or regulatory landscape on your genomic data?

6. End Users of the Proposed Results and Technology

- **a.** What national and international evidence supports the effectiveness of the proposed genomic technology in mitigating climate change?
- **b.** Please share national and/or international examples (if available) of adoption of your project's genomic solution.
- **c.** State the primary users of the technology developed in your project.
- **d.** Describe how the potential end users would have access to the technology and how it can be implemented.
- **e.** Describe the challenges/roadblocks/gaps that can impact the flow of genomic data from the ICT to end users

Appendix C – Public Data Management Example in the United States: Ag Data Commons

Ag Data Commons is the US Department of Agriculture's public research data catalogue and repository, operated by the National Agricultural Library. It provides open access to datasets created partially or entirely with Department of Agriculture funding and supports discovery across many agricultural domains rather than a single species or commodity. In practice, Ag Data Commons serves both as a repository and as a catalogue that can register externally hosted datasets while preserving a single discovery point for users. This allows project teams to deposit data directly when appropriate or link to domain repositories when size, format, or policy considerations make external hosting preferable [109]. Ag Data Commons is situated within a broader open data and public access program. The National Agricultural Library publishes contributor policies that define collection scope, required and optional metadata, licensing expectations, and conditions for record retention and versioning. For example, it recommends the use of standard open licenses to maximize legal interoperability. The library also publishes a data submission manual that walks contributors through submission and accessioning, which raises baseline metadata quality and reduces friction. All submitted datasets are reviewed by Ag Data Commons curators, who are metadata librarians. They verify completeness, licensing, and adherence to repository policies before approving public release. This human curation step ensures data integrity and consistency across the repository. Together, these steps codify scope, roles, and curation practices so records can be discovered and reused beyond the original project context [22], [110]. Metadata choices align with the US Data Catalog Vocabulary (DCAT-US), formerly known as the Project Open Data Metadata Schema [111], [112]. DCAT is a vocabulary from the World Wide Web Consortium that was originally developed in the context of US and United Kingdom federal government data catalogues, and is widely used across different sectors, including government, industry, research, and open data portals. DCAT defines classes and properties for describing datasets, their downloadable distributions, and catalogues, enabling consistent machine-readable metadata and cross-portal harvesting [111]. This alignment enables machine harvesting and cross-catalogue interoperability, and simplifies reuse of metadata in agency-wide discovery services. For a multi-project initiative, adopting a minimal profile that maps cleanly to DCAT-US or DCAT fields is an efficient way to future-proof cataloguing without prescribing heavy, bespoke schemas [111], [112].

Ag Data Commons also links to the Department of Agriculture's literature systems. PubAg serves as a discovery portal for research outputs that are authored and funded by the Department of Agriculture, while AGRICOLA indexes books and selected journal citations within the library's broader catalogue. A routine practice is to cross-reference articles and datasets by DOI so publications lead users to the underlying data and dataset records lead users to the publications. The same approach could work for the ICTs by making dataset DOIs and article citations required fields and by ensuring bidirectional links wherever possible [113]. Furthermore, the following operational features of Ag Data Commons could be easily translated into design principles for ICT data infrastructure:

- **Systematic use of persistent identifiers:** Assign DOIs to datasets and software and use ORCID for contributors to improve traceability and credit assignment.
- **A hybrid repository-plus-catalogue model:** Allow both direct deposits and registration of external records while providing a single discovery surface for users.

- **Clear public guidance:** Publish a contributor policy and a submission manual to create a predictable, high-quality process across heterogeneous projects.
- **Standardized metadata:** Align the cross-initiative metadata profile with DCAT-US or DCAT to ensure interoperability and make harvesting straightforward.
- **Simple, clear governance:** Place the data hub within a visible open data program that names roles, stewardship expectations, and refresh processes across the data lifecycle.

Two additional linkages round out the picture. Where teams have already deposited data in general repositories, such as Zenodo, Ag Data Commons registers those deposits rather than forcing relocation, as long as records meet the metadata profile. Likewise, when agency archives move or consolidate, the National Agricultural Library can serve as a durable home for discovery and access. Both patterns emphasize a pragmatic posture, which is to focus on a consistent catalogue view and on metadata that travels, not on centralizing every file in one place [22], [114]. There are also boundaries regarding what transfers directly. Ag Data Commons is a generalist agricultural repository and does not serve controlled-access human clinical genomics. For any ICT that touches human or sensitive health-related data, governance should follow the Global Alliance for Genomics and Health principles, with appropriate privacy controls and access committees, while still exposing non-sensitive discovery metadata. In addition, the legal context that shapes DCAT-US and US Department of Agriculture policy is not identical to Canadian federal and provincial frameworks. World Wide Web Consortium recommendations leading to data standards, such as DCAT-US and DCAT, could be reused for data interoperability, but governance language, hosting choices, and service levels should reflect Canadian institutions, funders, and Indigenous data stewardship requirements [111], [112], [115].

In summary, Ag Data Commons offers a workable model for a multi-domain, government-run agricultural data platform that combines a repository with a catalogue, publishes clear contributor guidance, relies on persistent identifiers, and aligns metadata with widely adopted standards. The most transferable actions for the ICTs are to adopt a lightweight persistent identifier policy for datasets, software, and people, to publish a minimal metadata profile mapped to DCAT-US or DCAT, to issue a concise contributor manual that matches that profile, and to operate a hub that can index both hosted deposits and externally archived records. Adopting these practices would build a foundation for interoperability across the ICTs' diverse data, from crops and livestock to soils and microbes, while lowering long-term curation costs [110], [111], [112].

Appendix D – Questionnaire for the Workshop Follow-Up Interviews

The goal of this questionnaire is to refine standardization needs and hurdles identified through the workshop discussion.

1. **Metadata Template**

   At the beginning of the workshop, we identified that 50% of data users do not have a metadata template in place that can be used by team members working with genomic data. In addition, developing and adopting unified metadata was considered the most important recommendation from the workshop audience.

   1. What do you consider the main reasons for not having a metadata template?
   2. How would the adoption of a standardized metadata template (within and across projects) impact your work? If you have a metadata template, what benefits have you seen?

2. **Cross-Project Data Practices**

   The implementation of cross-project data governance and management was also considered an important recommendation that needs to be prioritized when adopting standards for genomic data.

   1. What are the barriers to implementing unified data governance and management practices across different agricultural genomic projects?
   2. How would the adoption of unified practices impact your field (as a data user or industry)?
   3. What steps can be taken to establish effective cross-project data governance and management within Canada?
   4. What funding mechanisms or incentives would best support the adoption of genomic data standards in Canada, and what role should research funding agencies play in supporting cross-project implementation?

3. **Data Ownership and Sovereignty**

   During the workshop, data ownership and the importance of maintaining sovereignty over agricultural genomic information and repositories were highlighted as aspects that should be considered when implementing standards.

   1. Can you share any experiences or concerns related to data ownership and sovereignty that have influenced your approach to managing data?
   2. What issues of data ownership and sovereignty over agricultural genomic information and repositories should be addressed when implementing standards?
   3. How can researchers, the industry and funding agencies work together to ensure data ownership and sovereignty are maintained when implementing data standards?
   4. How can recognizing data ownership impact data quality and increase its potential for reuse?

CSA Group Research

In order to encourage the use of consensus-based standards solutions to promote safety and encourage innovation, CSA Group supports and conducts research in areas that address new or emerging industries, as well as topics and issues that impact a broad base of current and potential stakeholders. The output of our research programs will support the development of future standards solutions, provide interim guidance to industries on the development and adoption of new technologies, and help to demonstrate our on-going commitment to building a better, safer, more sustainable world.

CSA GROUP™

© 2026 Canadian Standards Association. All Rights Reserved.