Journeying towards best practice data management in biodiversity genomics

Running title

Biodiversity genomic data management

Authors

Natalie J. Forsdick*[1,2], Jana Wold*[3,2], Anton Angelo[4], François Bissey[5], Jamie Hart[5], Mitchell Head[6], Libby Liggins[7,2], Dinindu Senanayake[8], Tammy E. Steeves[3,2]

Affiliations

1 Manaaki Whenua – Landcare Research, New Zealand

2 Genomics Aotearoa, New Zealand

3 School of Biological Sciences, University of Canterbury, New Zealand

4 Library, University of Canterbury, New Zealand

5 Digital Services, University of Canterbury, New Zealand

6 Te Kotahi Research Institute, University of Waikato, New Zealand

7 School of Natural Sciences, Massey University, New Zealand

8 New Zealand eScience Infrastructure, New Zealand

* Co-first authors.

Corresponding author

NJF: forsdickn@landcareresearch.co.nz

Key words

CARE Principles for Indigenous Data Governance, data lifecycle, data management plans, digital sequence information, FAIR Guiding Principles, Indigenous data sovereignty

Manuscript version

preprintv2023.04.28

Abstract

Advances in sequencing technologies and declining costs are increasing the accessibility of large-scale biodiversity genomic datasets. To maximise the impact of these data, a careful, considered approach to data management is essential. However, challenges associated with the management of such datasets remain, exacerbated by uncertainty among the research community as to what constitutes best practices. As an interdisciplinary team with diverse data management experience, we recognise the growing need for guidance on comprehensive data management practices that minimise the risks of data loss, maximise efficiency for stand-alone projects, enhance opportunities for data reuse, facilitate Indigenous data sovereignty and uphold the FAIR and CARE Guiding Principles. Here, we describe four personas reflecting user experiences with data management to identify data management challenges across the biodiversity genomics research ecosystem. We then use these personas to demonstrate realistic considerations, compromises, and actions for biodiversity genomic data management. We also launch the Biodiversity Genomics Data Management Hub (https://genomicsaotearoa.github.io/data-management-resources/), containing tips, tricks and resources to support biodiversity genomics researchers, especially those new to data management, in their journey towards best practice. We aim to support the biodiversity genomics community in embedding data management throughout the research lifecycle to maximise research impact and outcomes.

Introduction

The field of biodiversity genomics has undergone a fast-paced transformation over the last decade. Once largely inaccessible for non-model organisms, advancements in sequencing technology have substantially reduced costs associated with generating these data, leading to significant increases in the types and volumes of genomic data. Today, biodiversity genomics is a highly dynamic research field that integrates methods pioneered in human health (e.g., genome-wide association studies; Ozaki et al., 2002), agricultural breeding programmes (e.g., inbreeding coefficients; Wright 1922), and principles from molecular ecology and evolution (e.g., identifying the genomic consequences of small population size; Khan et al. 2021; Liu et al. 2021; Duntsch et al. 2021; Robledo-Ruiz et al. 2022). The proliferation of data is being utilised to address an ever-expanding array of research questions and is a challenge for existing data management systems and research community practices.

To maximise the short- and long-term impacts of biodiversity genomic data, a considered and careful approach to data management is essential. Good data management practices (see Box 1) can benefit research teams and institutions, the research community, and wider society when biodiversity genomics data is used to address contemporary socio-environmental challenges. For research teams, the positive impacts of data management can be particularly pronounced for large and long-term projects where there is regular turnover of members and/or research roles are highly partitioned. Effective data management benefits research teams through ensuring efficient resource use (e.g., time, computational, financial), risk mitigation (e.g., data loss, misinterpretation, misuse), signalling credibility through data reproducibility (Baker, 2016; Eisner, 2018), and ease of data-sharing for enhanced collaboration (Lau et al., 2017; Möller et al., 2017; Riginos et al., 2020). For research institutes and/or funding organisations there may be legal obligations and long-term responsibilities (including social licence requirements) for them as custodians to maintain the integrity of research data. Furthermore, these information-rich biodiversity datasets have immense reuse value that can only be realised if the data-generating researchers/institutions undertake careful data management (Toczydlowski et al., 2021). These secondary use cases may diverge from the original purpose of data generation (Hoban et al., 2022; Leigh et al., 2021), and can provide additional valuable insights (e.g., Crandall et al., 2019), enhancing the value of these data to the research community and their potential impacts on society (e.g., Beninde et al., 2022; Exposito-Alonso et al., 2022).

Box 1. Best practices vs. good practices

Here we recognise there are different standards of data management. We acknowledge that achieving best practices is aspirational, and may not always be practicable within the constraints of a research project due to external factors (see section *Exploring biodiversity genomic data management challenges*). Instead, we encourage researchers to pursue 'good practices' as a stepping stone on the journey towards best practices.

Despite the availability of data management knowledge and resources, we acknowledge (and have lived experience with) the array of challenges inherent to the institutional frameworks in which we operate. These challenges may restrict the ability of research teams to adhere to best practices described herein. For example, the prevalence of short-term research contracts, combined with a 'publish or perish' mindset, may result in the deprioritisation of data management for some researchers. Nonetheless, even incremental improvements to data management by individuals, within their own capacity, should be encouraged and supported.

The incentives to implement data management practices are clear, and although there exists conceptual guidance on best practices within the broader scientific community (e.g., the FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al., 2016; and the CARE Principles for Indigenous data governance, Carroll et al., 2020, 2021), implementation remains challenging (Box 2). Contributing factors include the sheer volume of these information-rich datasets and the associated resource requirements (i.e., the time and financial costs of data curation, maintenance, and processing (Batley & Edwards, 2009; Chiang et al., 2011; Grigoriev et al., 2012; Schadt et al., 2010), as well as the inability of existing data standards, infrastructures, and repositories to keep pace with the needs of this research community (e.g., Crandall et al., 2023; Liggins et al., 2021). Best practices for biodiversity genomic data management are an active area of discussion among the biodiversity genomics community (Anderson & Hudson, 2020; Fadlelmola et al., 2021; Field et al., 2008; Liggins et al., 2021; Yilmaz et al., 2011). However, these initiatives can be easily missed by biodiversity genomics researchers because they are often disseminated as discipline-specfic outputs (e.g., publications, conference presentations, blogs) or institution-specific internal documents. Thus there are opportunities to centralise these existing resources. There are also benefits for research teams in extending their networks beyond the biodiversity genomics community to leverage the wealth of knowledge available across disciplines and institutes.

By necessity, biodiversity genomics brings together diverse teams with broad interests. We are a cross-institutional, interdisciplinary, multi-career stage collaborative team based in Aotearoa New Zealand, including biodiversity genomics researchers (NJF, JW, LL, TES), institutional and national eResearch and libraries staff (AA, FB, JH, DS), and those with broad interests in the inclusion of Indigenous perspectives pertaining to biodiversity genomic data (NJF, JW, MH, LL, TES). Our extensive experience includes: overseeing biodiversity genomic research projects, curating and managing biodiversity genomic datasets, developing project-specific data management plans (DMPs), and providing data management solutions to research teams. We have lived experience with the caveats of applying data management theory to real-life research situations.

Through this contribution we aim to provide support to biodiversity genomics researchers in incorporating data management within their daily research practices by:

- describing typical data management experiences of individuals across the research ecosystem;
- presenting ‘tips and tricks’ for documenting and managing genomic datasets and suggesting simple tools to support researchers in adhering to the FAIR and CARE Guiding Principles;
- collating resources such as templates and workflows for data management that can be readily adopted and/or adapted for wide usage in biodiversity genomics projects in the Biodiversity Genomics Data Management Hub (https://genomicsaotearoa.github.io/data-management-resources/).

We encourage researchers to view data management practices as behaviours intrinsic to the research process, and to adopt a mindset of adaptability to the various hurdles that may be encountered along the way. Through sharing these perspectives, we hope to support emerging researchers and the biodiversity genomics community more broadly on their data management journeys, and ultimately to amplify the real-world impacts of biodiversity genomics research.

Box 2. Ethical considerations for biodiversity genomic data management

The potential for data misuse (e.g., cherry-picking, data theft, unpermitted use, sharing, or misappropriation) is ever-present throughout the data lifecycle (Cragin et al., 2010). Data misuse is harmful to the integrity of the research, science, and innovation sector, and has important social implications due in part to an erosion of public trust in science (Laurie et al., 2014). Misuse can have direct negative impacts for participants, communities, research partners, and end-users. This harm can further extend to the research team, collaborators, and their institutes in the form of serious legal implications, reputational risk, and negative impacts on career trajectories. There are clear ethical processes for other aspects of research (such as regulatory bodies for human and animal ethics) but such ethical frameworks may not yet be established for the generation and storage of biodiversity genomic data (especially eDNA, plants, invertebrates, fungi). Data management is a tool researchers can use to mitigate this risk and some institutes and communities are well-versed in defining and implementing consistent and effective data management practices. However we recognise that there remain gaps between knowing and doing, with different groups positioned at different points on their data management journeys. Nonetheless, good data management minimises the risks of data misuse, loss, or theft, improves transparency, and ensures data FAIRness within established parameters specific to those data.

It also seeks to find balance between ‘Open Data’ and ‘Accessible Data’, the latter of which may be more appropriate for data pertaining to species and locations significant to Indigenous Peoples (e.g., Henson et al., 2021; Rayne et al., 2022). To facilitate Indigenous data sovereignty, data should be accompanied by metadata that includes details of appropriate permissions, which may include access restrictions. Local Contexts Notices, including Traditional Knowledge and Biocultural Labels, offer one such framework to support this (Anderson & Hudson 2020; Liggins et al., 2021).

Exploring biodiversity genomic data management challenges

Here we present user experience personas to describe data management needs for individuals in different career stages and roles. Using these personas, we aim to highlight some of the many important considerations associated with genomic data management. While we acknowledge that real life is not typically this tidy, we hope that researchers may see their own experiences reflected through some combination of these personas. The layers of challenges experienced by researchers may include the growing volume and types of genomic data and metadata, rapid technological and methodological advances, ensuring interoperability with metadata, and balancing openness and Indigenous data sovereignty.

Persona 1. A student new to biodiversity genomics

New PhD student Taylor Smith (Figure 1) has started a research project that will generate genomic data to inform conservation management for a culturally significant species (a recently described species of endemic lizard). Their project involves data collection and generation, analysis using the local compute infrastructure provided by their institute, and dissemination of results to end-users including conservation practitioners and local communities. They will be operating under a DMP adapted from the template used across their research team, and they have access to internal training and external support structures.

Their research team is in the process of developing a research manual that includes daily data management processes, along with on/offboarding procedures. Taylor is grateful for the supportive research environment, as they feel comfortable asking questions and sharing thoughts to help develop these processes. While their data is yet to be generated, being involved in these processes ensures they have a clear understanding of what will be involved in managing their data.

Taylor's main concerns are in ensuring their data management practices facilitate Indigenous data sovereignty and uphold the FAIR and CARE Guiding Principles during the active life-span of the project. As the project has a defined end-date, they also want to ensure that there is a framework in place to maintain these practices into the future. Communication around data management is primarily with their research team leader, Professor Nepia (Persona 3), who maintains trust-based relationships with the Indigenous tribes that have strong cultural ties to the focal species, and supported by the wider research team and eResearch and libraries staff.

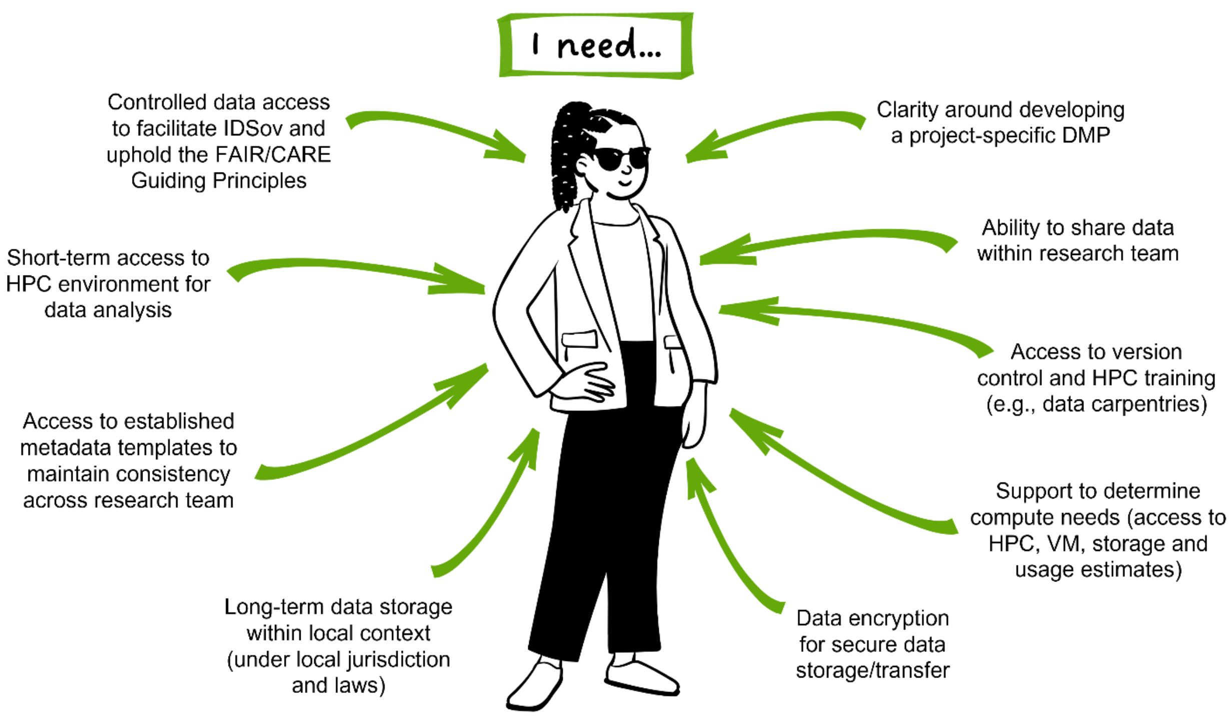

Figure 1. Examples of some typical data management needs and concerns that emerging researchers such as the persona of Taylor Smith are likely to have at the beginning of their data management journeys. DMP: Data Management Plan. HPC: High-performance compute. IDSov: Indigenous data sovereignty. VM: Virtual machine.

Persona 2. An early career researcher working collaboratively outside of academia

Dr Atsushi Sato (Fig. 2) is a postdoctoral researcher at a national research institute, and contributes to several large international biodiversity genomics collaborations (including with Professor Nepia, Persona 3). These projects vary in scale, longevity, and data management requirements. Each project Dr Sato is involved with has its own established DMP, so he must take care to ensure that the workflows he uses for each project align with the respective DMPs. Although he has some input in research planning and dissemination of results, his primary focus is on the analysis of large datasets, and specifically in incorporating environmental and climate data alongside genomic data. To do this, he relies on comprehensive and consistent metadata for each dataset.

He is experienced in biodiversity genomics, and is able to clearly report his data management needs to eResearch and libraries staff at his research institute. These needs predominantly relate to short-/mid-term storage and access, as the long-term storage of most of the datasets Dr Sato works with is the responsibility of researchers at other institutes. Dr Sato also seeks support from eResearch staff that deliver the national high-performance computing (HPC) infrastructure, where he can harness multithreading and parallel-processing for analysing these large datasets.

While Dr Sato's skills are in high demand, he has been persistently employed on precarious short-term contracts. He finds this stressful, and is constantly looking for new opportunities that may propel him towards his goal of attaining a permanent research position. These concerns impact his research priorities, as he perceives trade-offs between time spent on data management and that spent on data analysis that can produce results that contribute towards his publication record. From Dr Sato's perspective, data management is an onerous task.

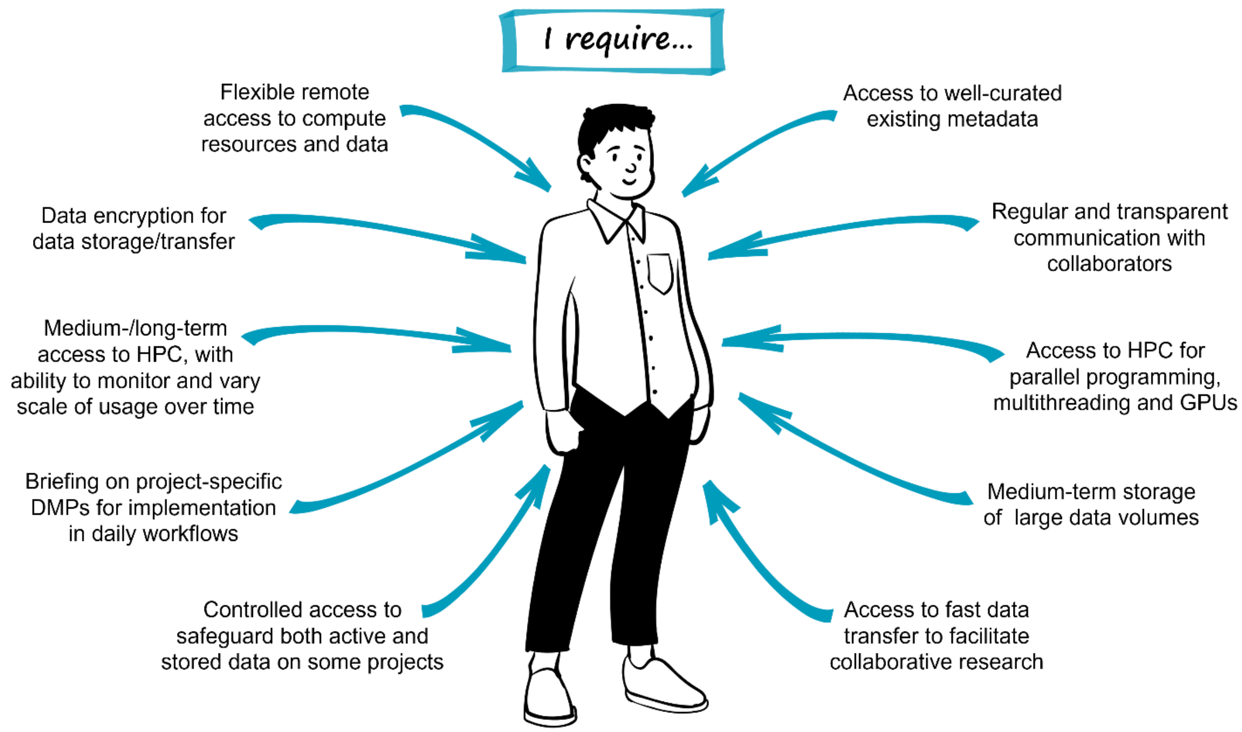

Figure 2. Examples of typical data management requirements experienced by researchers working in highly collaborative spaces, as exemplified by the persona of Dr Atsushi Sato. DMPs: Data Management Plans. HPC: High-performance computing. GPUs: Graphics processing units, often used to accelerate data processing.

Persona 3. A biodiversity genomics research team leader

Professor Tehara Nepia (Fig. 3) is a principal investigator at a university overseeing a conservation genomics research team including postgraduate students (including Taylor Smith, Persona 1), postdoctoral researchers, and research associates (including Dr Atsushi Sato, Persona 2). Her focus is on designing, facilitating, and disseminating research, and providing a supportive environment that produces highly-skilled emerging researchers well-equipped to contribute to the research, science, and innovation sector. Professor Nepia also places strong emphasis on building and maintaining trusted relationships with research partners, including Indigenous tribes. A substantial part of her role includes seeking and managing funding and resources (including compute and data storage) for the research team.

As the volume of data generated by Professor Nepia's team is continually expanding, there is a growing need to ensure a smooth transition of data (including metadata) between members of her research team. While Professor Nepia has a responsibility to meet institutional requirements, she is also committed to embedding data management practices that facilitate Indigenous data sovereignty and uphold the FAIR and CARE Guiding Principles. She is working towards a DMP template for use across all her research team's projects. To achieve this, Professor Nepia encourages open two-way communication with her research team to gain their perspectives of the needs and challenges associated with data management. She relies upon her research team to adhere to the DMPs, to support and encourage each other to do this, and to seek strategic advice from her when needed. Beyond the DMPs, Professor Nepia and her team co-develop research group guidelines that include data management practices to streamline team on/offboarding, allowing new members to quickly get up to speed, and providing clear expectations of data management for those departing.

She also engages with colleagues in similar situations nationally and internationally, including her disciplinary research community. Keeping abreast of evolving best practices in the biodiversity genomics research community and updating the research team's DMP template accordingly is an added pressure on Professor Nepia's limited time; she never feels completely up-to-date with the latest developments but understands she must be the one in the research team to lead data management practices even if she is only able to support 'good' versus 'best' practice (Box 1). To help with this burden, Professor Nepia prioritises building strong relationships with local eResearch and libraries staff (including Darryl, Persona 4) that are based on transparent, timely, bi-directional communication. Through knowledge-sharing, eResearch and libraries staff help her to understand local data management capacity and constraints, and gain the necessary understanding of the project-specific nuances that enable delivery of wrap-around solutions that support the needs of the research team now and into the future.

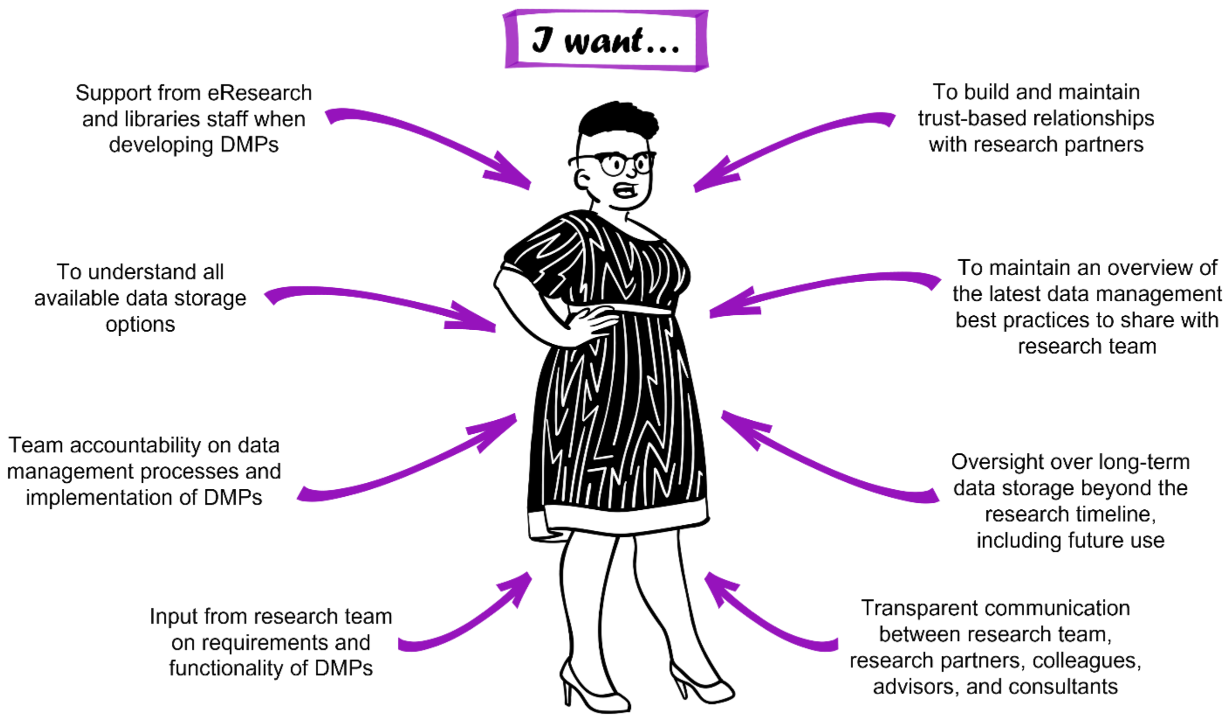

Figure 3. Examples of the types of support and level of oversight that research project leaders such as the persona of Professor Tehara Nepia may require when facilitating the development of consistent data management practices within their research teams. DMPs: Data Management Plans.

Persona 4. An eResearch staff member

Darryl Baker (Fig. 4) is an eResearch Manager at a university, and provides eResearch support to numerous research projects across all disciplines and departments, including providing advice and services relating to compute and data storage facilities for biodiversity genomic data. Darryl manages the resource that is the institutional compute and storage facilities allocated to research. He keeps up to date with research-focused technologies, consults with research teams, and mentors researchers on the use of the available research systems. In the last four years the storage facility of the institution has reached peak capacity, requiring careful resource management. Darryl seeks budget approval to expand the current on-premise storage facility. Based on quotes provided by vendors, purchasing additional storage infrastructure proves to be expensive. Further, it would only provide a short-term fix as the institution's research data is predicted to exceed the storage limit within five years.

Recently, Professor Nepia (Persona 3) reached out to Darryl for eResearch services and support for her biodiversity genomics research team. Professor Nepia's team generates a number of projects, with rapidly increasing data management needs over the last 10 years. Darryl meets with one of Professor Nepia's research students, Taylor Smith (Persona 1), to understand the eResearch needs of an upcoming project about a new species of lizard. In a face-to-face meeting, he gathers information about the data being produced. Early indications are that this project will generate vast amounts of data and function under a DMP. Darryl wishes to understand the project-specific needs in order to advise on appropriate storage and computing solutions that will facilitate Indigenous data sovereignty and uphold the FAIR and CARE Guiding Principles. Darryl holds a clear understanding of the constraints arising from the institutional infrastructure, and the responsibilities of the researcher under national and institutional legislation. Through conversations with researchers and research teams, Darryl can gain a clear vision of what they are trying to achieve within these constraints, and provide advice and solutions to overcome data management pain points that may arise.

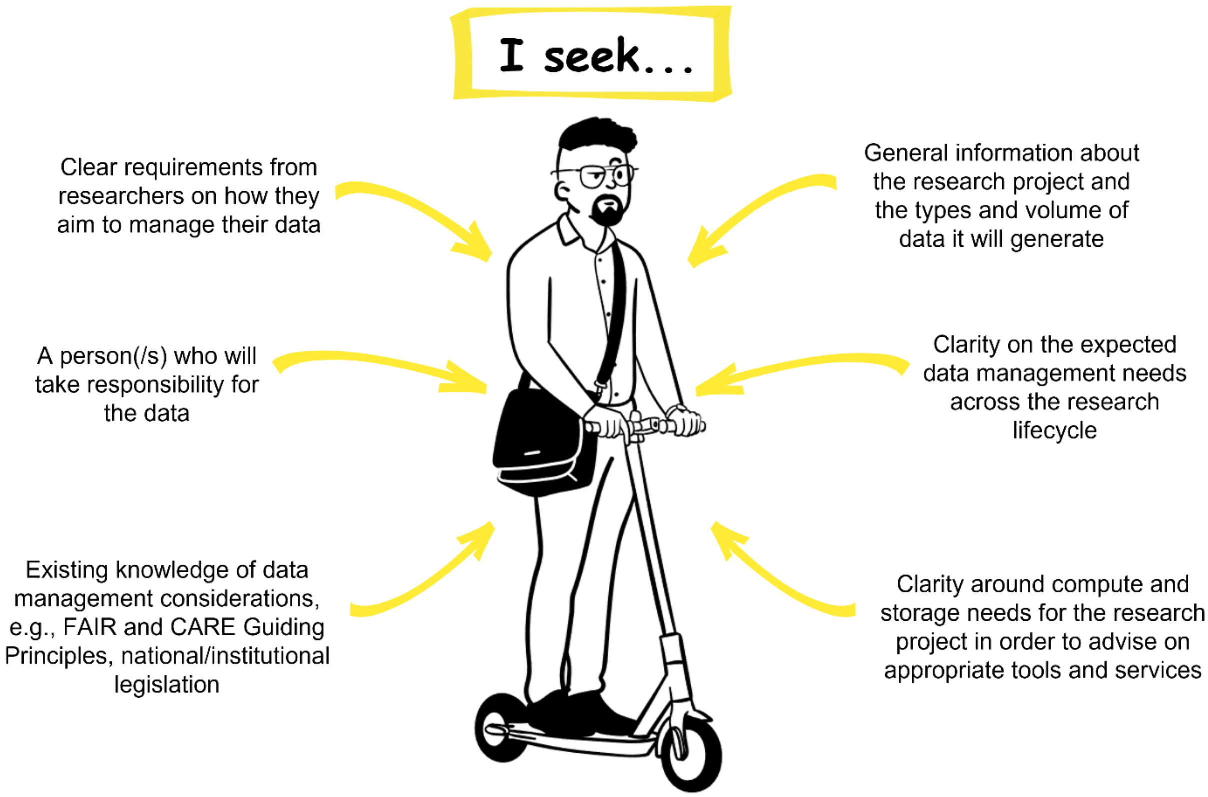

Figure 4. Examples of typical needs of eResearch and libraries staff such as the persona of Darryl Baker in the development and delivery of specialised data management solutions for researchers and research teams.

Addressing the challenges

Following the description of these personas, we identified key data management questions that researchers across the biodiversity genomics research ecosystem may have, and propose solutions to support good data management practices (Fig. 5). As every situation is different, we recognise that not all solutions will be immediately adaptable to specific challenges, but may spark ideas. Here we provide discussion of some potential solutions to these identified challenges, and supporting resources to implement effective data management practices.

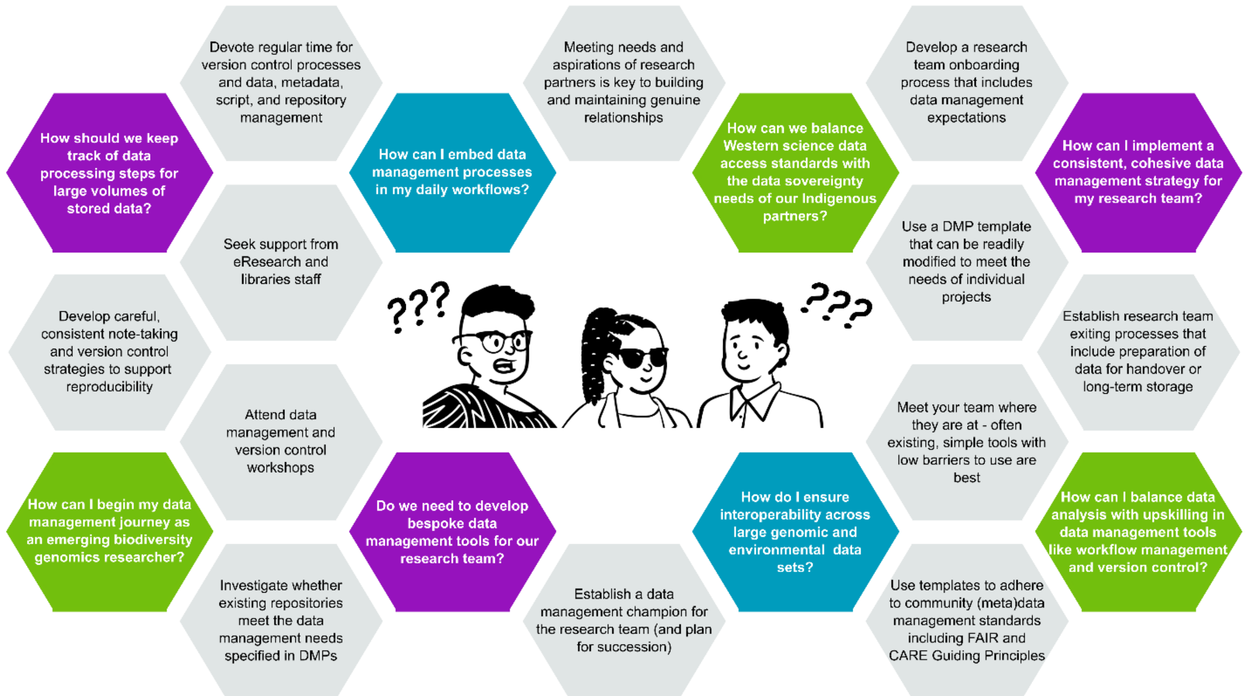

Figure 5. Key data management questions (coloured hexagons) that biodiversity genomic researchers and teams may have, along with potential (non-exhaustive) solutions (light grey hexagons) to support them during their data management journeys. Colours of the question hexagons are used to denote their relevance to the personas described above, though we note that different personas may share common questions, and that solutions may address multiple challenges.

1. Resources to support researchers in implementing effective data management

To reduce the frustration often experienced by researchers on their journey towards best practices in data management, we have established the Biodiversity Genomics Data Management Hub (https://genomicsaotearoa.github.io/data-management-resources/) where we connect the challenges described in the personas to modules that provide topic-specific tips, tricks, and resources, including from beyond the traditional biodiversity genomics literature. Module content draws on the diversity of our experiences and knowledge, with topics including: ‘Hot, warm, and cold data storage’, ‘Data Management Plans in practice’, and ‘Helping eResearch staff help you'. These tips and tricks are largely hard-won through the trials and tribulations experienced during our personal research journeys. We intend for the Hub to be a living resource that evolves over time, incorporating new tools and practices as these come to light. We welcome suggestions of additional module topics, along with contributions of the latest resources. We envision that the Hub will be of special interest for emerging researchers, and will be useful as a teaching resource, instilling data management practices as part of daily workflows from the beginning of the research journey. The Hub may also provide an opportunity for those with an interest in data management outside of the genomics space to have the opportunity to peek 'through the looking glass' and gain insight into the similarities and differences with their own fields.

In assembling resources for the Hub to address challenges across personas, three overarching actions stood out as immediately accessible steps toward best practices for the biodiversity genomics community. Here, we elaborate on these.

2. Develop Data Management Plans

Biodiversity genomic data management tends to come into focus at the end rather than throughout the research lifecycle. Many journals that publish biodiversity genomic research have open data policies (e.g., the Joint Data Archiving Policy), and this may be the first instance at which researchers are required to demonstrate data management. Indeed, genomics broadly appears immature compared with other disciplines in terms of data management. For example, DMPs are often perceived as 'nice to have' but are not yet widely required. However, when working with the large volumes of data produced via genomic sequencing, and/or in research teams distributed across multiple institutions, data management can quickly degenerate leaving the data, researchers, and research partners vulnerable (Box 2). Further, DMPs are one tool among many that will be required to achieve the benefit-sharing goals pertaining to genomic data as described in the Kunming-Montreal Global Biodiversity Framework (Decisions 15/4 and 15/9, https://www.cbd.int/decisions/cop/?m=cop-15).

DMPs are key tools for mitigating the risks of data loss and misuse. Where they do not already exist, we anticipate a widespread shift towards the establishment of data management policies within institutions and by research funding organisations (including the requirement of DMPs in research funding applications) in the near future (Bloemers & Montesanti, 2020; Fadlelmola et al., 2021; Jorgenson et al., 2021). Indeed, the primary research funding body in Aotearoa New Zealand, the Ministry of Business, Innovation and Employment, is shifting towards an open research policy (https://www.mbie.govt.nz/science-and-technology/science-and-innovation/agencies-policies-and-budget-initiatives/open-research-policy/) as many of its contemporaries have done (e.g., the Australian Research Council, the European Research Council, the National Institutes of Health), which may come to include a requirement for DMPs. We foresee that some of the challenges associated with requirements to provide DMPs during funding applications will be in ensuring cohesive frameworks for the development of DMPs that are fit for purpose, and more broadly in the development and maintenance of trusted data repositories at scale (Lin et al. 2020).

The inclusion of an approval and/or compliance pathway may be recommended to ensure that DMPs lead to meaningful actions in the improvement of data management in biodiversity genomics rather than simple ‘box-ticking’ or thought exercises. Specifically, approval pathways would require consideration of the DMP during the funding application process to determine whether it is fit for purpose. In comparison, a compliance pathway requires researchers to demonstrate that data management actions have been carried out in accordance with the DMP provided. DMP approval and compliance with regard to the FAIR Guiding Principles would require consideration by external assessment panels with discipline-specific knowledge and expertise. For data and metadata associated with species or locations significant to Indigenous Peoples (see Box 2), decisions around auditing and assessment of DMPs in relation to the CARE Guiding Principles can only be made by the associated Indigenous Peoples. Indigenous leadership will be essential in the co-development of any such systems, with one important consideration being ensuring that DMPs are responsive to current concerns while remaining flexible for the future. Indeed, there is unlikely to be a ‘one size fits all’ solution for culturally significant data.

While compliance is one method of ensuring that data management actions are implemented, research projects tend to change course over time, and a DMP designed during the planning stage may not provide the flexibility required to meet changing data needs later in the research lifecycle. Rather than using approvals or compliance processes to ensure appropriate data management actions are taken, a more appropriate approach could be to recognise a DMP as a live document throughout the research process, allowing for updates as the project changes. In this scenario, version control methods should be used to track changes throughout the project. During any process of revision of the DMP, it will be important to maintain regular and transparent communication with relevant research partners whenever changes are being considered, to ensure that changes are both fit for purpose, while continuing to accommodate the needs and interests of all parties. At the end of the project, the research team could complete a self-reflective retrospective process, identifying which aspects went according to plan, where needs changed over time, and whether there were any limitations or challenges due to institutional or infrastructure constraints. This could help researchers to better understand the capabilities and capacities of their teams and systems, and inform future research design that includes DMP development. Further, by feeding back the learnings derived through this retrospective to associated eResearch and libraries staff will help to close the loop.

3. Seek support from eResearch and libraries staff

We challenge researchers to look beyond their immediate research community for assistance - help may be closer at hand than expected. Here we highlight the benefits of engaging with eResearch and libraries staff within or beyond your institute from an early stage in the research lifecycle. These professional staff are a supporting network holding knowledge and expertise in crafting solutions to data management challenges (Andrikopoulou et al., 2022). Researchers benefit from developing these relationships with staff who cultivate institutional knowledge and solutions that may not be captured in the traditional or domain-specific scientific literature. eResearch and libraries staff can provide guidance and targeted support in the co-development of project-specific data management strategies that take into account institutional operating requirements and the capacity and capability of existing infrastructure, and in incorporating data management practices into day-to-day research workflows.

eResearch and libraries staff may at times be overlooked due to the frequent tangible and intangible siloing of disciplines, resulting in researchers being unaware of how these staff can provide support, and unclear as to what their mandates are, with eResearch and libraries staff consequently unaware of the data management needs and challenges experienced by research teams. Further, eResearch and libraries staff are often spread thinly across institutions, with high demand for their services but limited capacity to provide much-needed support. As such, building channels of communication between research teams and support staff is key, and both parties must be willing to come to the table to share and learn from one another.

Developing strong working relationships requires reciprocity, with an emphasis on mutual benefit (which may include academic acknowledgement) and respect for expertise on both sides. eResearch and libraries staff often require knowledge of the research context and learned experiences from researchers so they can provide and/or procure the necessary services and support, and researchers can also endeavour to engage with the technicalities and concepts necessary for full and fruitful discussions. We recommend that researchers meet early and often with eResearch and libraries staff to discuss their data management needs. Investing in these relationships ultimately means that researchers will get the wrap-around support they require, and eResearch and libraries staff will be kept appraised of their changing needs, facilitating the development of future-focussed solutions.

4. Establish a research data management culture in your team

It is vital to ensure the continuity of data management throughout the research lifecycle. We strongly encourage researchers to step up and take an active leadership role in situations where there is an absence of clear and consistent guidelines. However, data management is most effective when pursued as a team, with a consistent and cohesive plan and division of labour. A little effort early in the process can go a long way, and so we recommend that research teams develop clear documentation around on/offboarding procedures and daily data management practices. This will streamline the process of joining the team, provide guidance on the options for and constraints around data transfer, storage, and access, and a clear pathway to follow when departing that may include ongoing access to data, or the packaging of data and metadata for long-term storage.

To ensure consistency despite the potential for frequent turnover within the team, we suggest that research teams establish a data management champion to oversee the onboarding and training of new members and ensure the implementation of consistent data management practices across the research team. While anyone can take on this transferable role, a data management champion will ideally have a mid- to long-term position within the research team, hold a deep understanding of the unique characteristics of each research project, and have the necessary level of autonomy to operate independently as a leader in this role. Succession planning for this role will be essential to ensure consistency and continuity. This person can also operate as a conduit between the research team and eResearch and libraries staff, and so excellent people skills will be advantageous. By engaging regularly and often with their institute’s support structures, they can ensure that eResearch and libraries staff are kept up to date with the changing needs of the team, and ensure access to the latest services and support.

Continuing the data management journey

Here we have presented tips and tricks to support biodiversity genomics researchers in the development of good data management practices, though we acknowledge that any level of data management is better than none. Data management is a journey, and we are all on an aspirational path striving towards best practice. We trust our contribtion will be a helpful guide for researchers new to biodiversity genomics, and a useful prompt for existing researchers to embed good data management practices into their daily research routines.

Glossary

- Accessible data. Data accessible under well-defined conditions, as per the FAIR Guiding Principles (Mons et al., 2017; Wilkinson et al., 2016).
- CARE Principles for Indigenous Data Governance. Designed to complement the FAIR Guiding Principles, these people- and purpose-oriented principles and supporting concepts (Collective benefit, Authority to control, Responsibility, Ethics) reflect the crucial role of data in advancing innovation, governance, and self-determination among Indigenous Peoples (Carroll et al. 2020; 2021). https://www.gida-global.org/care.
- Data lifecycle. The steps in the research process specifically pertaining to data, from planning, collection and generation, analysis and collaboration, evaluation, storage, dissemination, access, and reuse, which can contribute to the planning for new data generation. The data and research lifecycles are distinct but interrelated.
- Data management. The processes and practices associated with the documentation and storage of and access to data and associated metadata throughout the research lifecycle.
- DMP. Data management plan. A document describing the data that will be generated during a research project, and how it will be used, accessed, and stored during the research lifecycle. Also known as a data management and sharing plan, though in our definition of data management, data sharing is inherently included in data access.

- eResearch. The use of digital tools and techniques to advance research.

- eResearch and libraries staff. A broad group that includes research software engineers, research infrastructure developers, data scientists, data stewards, and other professional services staff that deliver library, IT, bioinformatics, and high-performance compute support.

- FAIR Guiding Principles. Guidelines for scientific data management and stewardship intended to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets (Wilkinson et al. 2016). https://www.go-fair.org/fairprinciples/

- Indigenous data. The tangible and/or intangible cultural materials, belongings, knowledge, digital data, and information about Indigenous Peoples or that to which they relate (Lovett et al., 2019; Rainie et al., 2019).

- Indigenous data sovereignty. The expression of a legitimate right of Indigenous Peoples to control the access, the collection, ownership, application and governance of their own data, knowledge, and/or information that derives from unique cultural histories, expressions, practices, and contexts (https://localcontexts.org/indigenous-data-sovereignty/).

- Metadata. Data that provides information about other data. For biodiversity genomic data, metadata can provide information regarding context (e.g., taxonomic, spatial, temporal, and associated permissions) as well as used technologies/methodologies.

- Open data. Data anyone can use and share, typically openly accessible and with an open licence.

- Research lifecycle. The steps in the process of scientific research from inception (research planning, design, and funding) to completion (dissemination of results and real-world impact), which often leads back to development of new related projects. The research and data lifecycles are distinct but interrelated.
- VM: Virtual machine. A software-based computer system emulating that of a different physical machine, often used to run a different operating system than that of the primary system of the physical computer

Acknowledgements

We wish to thank the following people for their thoughtful advice, insights, and friendly feedback during the development of this project: Mik Black, Thomas Buckley, Eric D. Crandall, Manpreet Dhami, Tom Etherington, Leanne Elder, Stephanie Galla, Tipene Merritt and the University of Canterbury (UC) eResearch Co-Design Group, David Medyckyj-Scott, Nick Spencer, Matt Stott, and the UC ConSERTeam.

Author Contributions

NF, JW and TES conceived the research. All authors provided input into the research direction and contributed through robust discussion towards the development of the manuscript and the Biodiversity Genomic Data Management Hub. JH provided illustrations. NF and JW wrote the first draft of the manuscript, and led the writing of subsequent drafts. All authors provided feedback and approved the final manuscript.

Benefit-Sharing Statement

Benefits Generated: A cross-institutional, interdisciplinary research collaboration was developed with all collaborators included as co-authors. Benefits from this collaboration accrue through the provision of the Biodiversity Genomic Data Management Hub, which is shared as a publicly available web resource to support biodiversity genomics researchers in improving data management practices across the data lifecycle. This research is timely given predicted changes in research funding requirements to include Data Management Plans.

Data Accessibility Statement

No data was produced or analysed in the development of this manuscript.

References

Anderson, J., & Hudson, M. (2020). The Biocultural Labels Initiative: Supporting Indigenous rights in data derived from genetic resources. *Biodiversity Information Science and Standards*, *4*, e59230. https://doi.org/10.3897/biss.4.59230

Andrikopoulou, A., Rowley, J., & Walton, G. (2022). Research Data Management (RDM) and the Evolving Identity of Academic Libraries and Librarians: A Literature Review. *New Review of Academic Librarianship*, *28*(4), 349–365. https://doi.org/10.1080/13614533.2021.1964549

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), Article 7604. https://doi.org/10.1038/533452a

Batley, J., & Edwards, D. (2009). Genome sequence data: Management, storage, and visualization. *BioTechniques*, *46*(5), 333–336. https://doi.org/10.2144/000113134

Beninde, J., Toffelmier, E., & Shaffer, H. B. (2022). A brief history of population genetic research in California and an evaluation of its utility for conservation decision-making. *Journal of Heredity*, *113*(6), 604–614. https://doi.org/10.1093/jhered/esac049

Bloemers, M., & Montesanti, A. (2020). The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices. *Data Intelligence*, *2*(1–2), 171–180. https://doi.org/10.1162/dint_a_00039

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1), Article 1. https://doi.org/10.5334/dsj-2020-043

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(1), Article 1. https://doi.org/10.1038/s41597-021-00892-0

Chiang, G.-T., Clapham, P., Qi, G., Sale, K., & Coates, G. (2011). Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics*, *12*(1), 361. https://doi.org/10.1186/1471-2105-12-361

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Crandall, E. D., Riginos, C., Bird, C. E., Liggins, L., Treml, E., Beger, M., Barber, P. H., Connolly, S. R., Cowman, P. F., DiBattista, J. D., Eble, J. A., Magnuson, S. F., Horne, J. B., Kochzius, M., Lessios, H. A., Liu, S. Y. V., Ludt, W. B., Madduppa, H., Pandolfi, J. M., … Gaither, M. R. (2019). The molecular biogeography of the Indo-Pacific: Testing hypotheses with multispecies genetic patterns. *Global Ecology and Biogeography*, *28*(7), 943–960. https://doi.org/10.1111/geb.12905

Crandall, E. D., Toczydlowski, R. H., Liggins, L., Holmes, A. E., Ghoojaei, M., Gaither, M. R., Wham, B. E., Pritt, A. L., Noble, C., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Delgado, A., Farrell, E., Himmelsbach, N., Queeno, S. R., Trinh, T., Weyand, C., … Toonen, R. J. (2023). Importance of timely metadata curation to the global surveillance of genetic diversity. *Conservation Biology*, *00*(e14061). https://doi.org/10.1111/cobi.14061

Duntsch, L., Whibley, A., Brekke, P., Ewen, J. G., & Santure, A. W. (2021). Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Molecular Ecology*, *30*(23), 6006–6020. https://doi.org/10.1111/mec.16068

Eisner, D. A. (2018). Reproducibility of science: Fraud, impact factors and carelessness. *Journal of Molecular and Cellular Cardiology*, *114*, 364–368. https://doi.org/10.1016/j.yjmcc.2017.10.009

Exposito-Alonso, M., Booker, T. R., Czech, L., Gillespie, L., Hateley, S., Kyriazis, C. C., Lang, P. L. M., Leventhal, L., Nogues-Bravo, D., Pagowski, V., Ruffley, M., Spence, J. P., Toro Arana, S. E., Weiß, C. L., & Zess, E. (2022). Genetic diversity loss in the Anthropocene. *Science*, *377*(6613), 1431–1435. https://doi.org/10.1126/science.abn5642

Fadlelmola, F. M., Zass, L., Chaouch, M., Samtal, C., Ras, V., Kumuthini, J., Panji, S., & Mulder, N. (2021). Data Management Plans in the genomics research revolution of Africa: Challenges and recommendations. *Journal of Biomedical Informatics*, *122*, 103900. https://doi.org/10.1016/j.jbi.2021.103900

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., … Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, *26*(5), Article 5. https://doi.org/10.1038/nbt1360

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R. A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D., & Dubchak, I. (2012). The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, *40*(D1), D26–D32. https://doi.org/10.1093/nar/gkr947

Henson, L., Balkenhol, N., Gustas, R., Adams, M., Walkus, J., Housty, W., Stronen, A., Moody, J., Service, C., Reece, D., vonHoldt, B., McKechnie, I., Koop, B., & Darimont, C. (2021). Convergent geographic patterns between grizzly bear population genetic structure and

Indigenous language groups in coastal British Columbia, Canada. *Ecology and Society*, *26*(3). https://doi.org/10.5751/ES-12443-260307

Hoban, S., Archer, F. I., Bertola, L. D., Bragg, J. G., Breed, M. F., Bruford, M. W., Coleman, M. A., Ekblom, R., Funk, W. C., Grueber, C. E., Hand, B. K., Jaffé, R., Jensen, E., Johnson, J. S., Kershaw, F., Liggins, L., MacDonald, A. J., Mergeay, J., Miller, J. M., … Hunter, M. E. (2022). Global genetic diversity status and trends: Towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. *Biological Reviews*, *97*(4), 1511–1538. https://doi.org/10.1111/brv.12852

Jorgenson, L. A., Wolinetz, C. D., & Collins, F. S. (2021). Incentivizing a New Culture of Data Stewardship: The NIH Policy for Data Management and Sharing. *JAMA*, *326*(22), 2259–2260. https://doi.org/10.1001/jama.2021.20489

Khan, A., Patel, K., Shukla, H., Viswanathan, A., van der Valk, T., Borthakur, U., Nigam, P., Zachariah, A., Jhala, Y. V., Kardos, M., & Ramakrishnan, U. (2021). Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. *Proceedings of the National Academy of Sciences*, *118*(49), e2023018118. https://doi.org/10.1073/pnas.2023018118

Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., Bajcic, D., Radenkovic, J., Mladenovic, V., Krstanovic, D., Arsenijevic, V., Klisic, D., Mitrovic, M., Bogicevic, I., Kural, D., … Seven Bridges CGC Team. (2017). The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Research*, *77*(21), e3–e6. https://doi.org/10.1158/0008-5472.CAN-17-0387

Laurie, G., Jones, K. H., Stevens, L., & Dobbs, C. (2014). *A Review of Evidence Relating to Harm Resulting from Uses of Health and Biomedical Data* (p. 210). Nuffield Council on Bioethics. https://www.pure.ed.ac.uk/ws/portalfiles/portal/19402878/Review_of_Evidence_Relating_to_Harms_Resulting_from_Uses_of_Health_and_Biomedical_Data_FINAL.pdf

Leigh, D. M., van Rees, C. B., Millette, K. L., Breed, M. F., Schmidt, C., Bertola, L. D., Hand, B. K., Hunter, M. E., Jensen, E. L., Kershaw, F., Liggins, L., Luikart, G., Manel, S., Mergeay, J., Miller, J. M., Segelbacher, G., Hoban, S., & Paz-Vinas, I. (2021). Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics*, *22*(12), Article 12. https://doi.org/10.1038/s41576-021-00394-0

Liggins, L., Hudson, M., & Anderson, J. (2021). Creating space for Indigenous perspectives on access and benefit-sharing: Encouraging researcher use of the Local Contexts Notices. *Molecular Ecology*, *30*(11), 2477–2482. https://doi.org/10.1111/mec.15918

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L’Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), Article 1. https://doi.org/10.1038/s41597-020-0486-7

Liu, L., Bosse, M., Megens, H.-J., de Visser, M., A. M. Groenen, M., & Madsen, O. (2021). Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. *Evolutionary Applications*, *14*(3), 710–720. https://doi.org/10.1111/eva.13150

Lovett, R., Lee, V., Kukutai, T., Cormack, D., Rainie, S. C., & Walker, J. (2019). Good data practices for Indigenous data sovereignty and governance. In *Good data* (pp. 26–36). Institute of Network Cultures Inc.

Möller, S., Prescott, S. W., Wirzenius, L., Reinholdtsen, P., Chapman, B., Prins, P., Soiland-Reyes, S., Klötzl, F., Bagnacani, A., Kalaš, M., Tille, A., & Crusoe, M. R. (2017). Robust

Cross-Platform Workflows: How Technical and Scientific Communities Collaborate to Develop, Test and Share Best Practices for Data Analysis. *Data Science and Engineering*, *2*(3), 232–244. https://doi.org/10.1007/s41019-017-0050-4

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, *32*(4), Article 4. https://doi.org/10.1038/ng1047

Rainie, S. C., Kukutai, T., Walter, M., Figueroa-Rodríguez, O. L., Walker, J., & Axelsson, P. (2019). Indigenous data sovereignty. In *The State of Open Data: Histories and Horizons* (pp. 300–319). African Minds and International Development Research Centre.

Rayne, A., Blair, S., Dale, M., Flack, B., Hollows, J., Moraga, R., Parata, R. N., Rupene, M., Tamati-Elliffe, P., Wehi, P. M., Wylie, M. J., & Steeves, T. E. (2022). Weaving place-based knowledge for culturally significant species in the age of genomics: Looking to the past to navigate the future. *Evolutionary Applications*, *15*(5), 751–772. https://doi.org/10.1111/eva.13367

Riginos, C., Crandall, E. D., Liggins, L., Gaither, M. R., Ewing, R. B., Meyer, C., Andrews, K. R., Euclide, P. T., Titus, B. M., Therkildsen, N. O., Salces-Castellano, A., Stewart, L. C., Toonen, R. J., & Deck, J. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources*, *20*(6), 1458–1469. https://doi.org/10.1111/1755-0998.13269

Robledo-Ruiz, D. A., Gan, H. M., Kaur, P., Dudchenko, O., Weisz, D., Khan, R., Lieberman Aiden, E., Osipova, E., Hiller, M., Morales, H. E., Magrath, M. J. L., Clarke, R. H., Sunnucks, P., & Pavlova, A. (2022). Chromosome-length genome assembly and linkage map of a critically endangered Australian bird: The helmeted honeyeater. *GigaScience*, *11*, giac025. https://doi.org/10.1093/gigascience/giac025

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, *11*(9), Article 9. https://doi.org/10.1038/nrg2857

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger, B., López, A., & Rehwald, S. (2020). When are researchers willing to share their data? – Impacts of values and uncertainty on open data in academia. *PLOS ONE*, *15*(7), e0234172. https://doi.org/10.1371/journal.pone.0234172

Toczydlowski, R. H., Liggins, L., Gaither, M. R., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Davis, B., Delgado, A., Farrell, E., Ghoojaei, M., Himmelsbach, N., Holmes, A. E., Queeno, S. R., Trinh, T., Weyand, C. A., Bradburd, G. S., Riginos, C., Toonen, R. J., & Crandall, E. D. (2021). Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences*, *118*(34), e2107934118. https://doi.org/10.1073/pnas.2107934118

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, *56*(645), 330–338. https://doi.org/10.1086/279872

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), Article 5. https://doi.org/10.1038/nbt.1823