FAIRification of Genomic Annotations Working Group

Case Statement

1. Charter

Motivation: Since the completion of the Human Genome Project, we have witnessed an explosion of datasets annotating particular locations along reference DNA sequences with e.g. aspects of functional processes. Tremendous amounts of research funding have been provided to large and small projects that have generated genomic annotation datasets. Many of these datasets relate to human genomes, but an increasing amount of such data is also generated for other model organisms – and with the recent surge of biodiversity projects – for a range of species spanning the whole tree of life. Unfortunately, researchers who want to make use of such data face practical challenges in discovering and reusing datasets relevant to their research. While many repositories and data distribution solutions exist, the metadata is often poorly aligned with best practices for Findable, Accessible, Interoperable and Reusable (FAIR) research data[1]. Even in cases where proper metadata exists, relevant solutions typically provide metadata according to different metadata models and APIs.

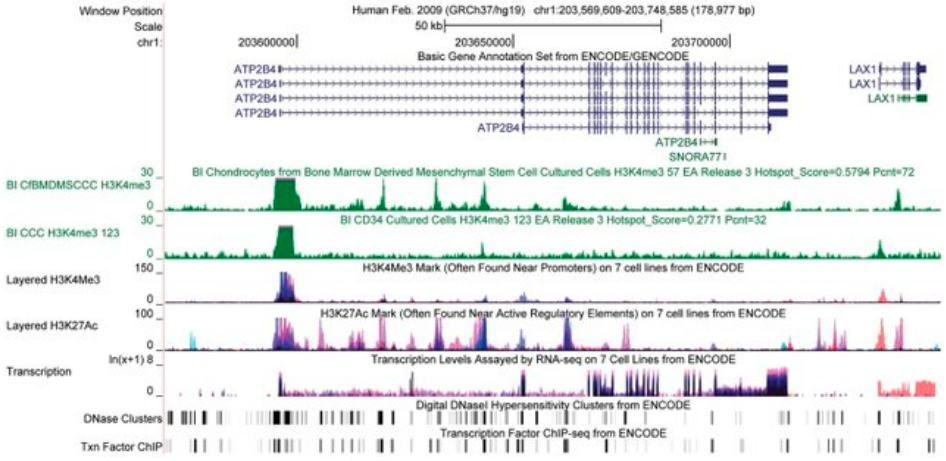

Figure 1: Example of genomic annotations (tracks) from the ENCODE consortium imported into the UCSC Genome Browser. From Rosenbloom et. al. (2011). "ENCODE whole-genome data in the UCSC genome browser: update 2012." Nucleic acids research. 40. D912-7. Licence: CC BY-NC 3.0

[1] e.g.; Xue, Bingjie, et al. "Opportunities and challenges in sharing and reusing genomic interval data." Frontiers in Genetics 14 (2023): 1155809; Gonçalves, Rafael S., and Mark A. Musen. "The variable quality of metadata about biological samples used in biomedical experiments." Scientific data 6.1 (2019): 1-15.

Genomic annotation data: The FAIRification of Genomic Annotations Working Group (FGA-WG) focuses on the challenges of harmonising metadata and software solutions to improve the discovery and reuse of publicly available “genomic annotation” data (see Figure 1). Genomic annotations, sometimes referred to as “genomic tracks[2]”, refer to data files that annotate reference sequence positions and can be visualised in genome browsers, such as the UCSC Genome Browser and Ensembl, or analysed often non-visually using computational tools that span the domains of life science (e.g., statistical colocalization analyses). Genomic annotations are often routinely generated by computational workflows in larger datasets; in the context of software- or method-oriented publications; or as the result of manual curation processes. The types of data include condensed summaries of experiment outputs as well as predictive and descriptive annotation of DNA reference sequence positions.

Deliverables: The FGA-WG establishes a minimal metadata schema based on the harmonisation of existing schemas and recommendations, working to gradually improve this schema based on the interrogation of a set of use cases. To foster the availability of metadata that adheres to the schema, the FGA-WG will develop practical recommendations to build scalable and maintainable metadata transformation pipelines to "FAIRify" and unify metadata from multiple data sources. Recommendations will be developed for publishing and registering the harmonised metadata – with persistent and globally unique identifiers – such that the metadata can be easily harvested by search engines and other services that improve the discovery and reuse of the metadata (and associated datasets). Lastly, the FGA-WG will develop a harmonised API for the use of search and discovery services by downstream users and tools.

The three main use cases to be considered are the following:

> **1. Biomedical analysis:** Discovery of genomic annotations through harmonised metadata to improve the availability and scope of datasets for analytical applications, including methodology-oriented software tools and AI/ML efforts. The focus area will be human biomedical research, but this use case could also include support for research on other species insofar as this aligns with the interests of the FGA-WG contributors.
>
> **2. Biodiversity genome annotation:** Establish genome annotations for biodiversity assemblies as FAIR objects with improved metadata and integration with relevant infrastructure and tools, including both the manual and automated annotation of genome assemblies.
>
> **3. Track Hub infrastructure:** Enhance genome browser-related infrastructure for track hubs with harmonised metadata and related services, to simplify the process of generating, hosting, and registering metadata-enriched track hubs; improve discovery of track hub-hosted data at the level of individual tracks; and facilitate the integration of metadata-enriched track hubs with genome browsers and other analysis tools.

Community building: The FGA-WG aims to build a broad community of individuals and groups with an interest in data integration related to genomic annotations, encompassing data producers, domain experts, tool/service developers, FAIR/RDM specialists, ethics and ELSI expertise, and analytical end users. The long-term goal is to build a sustainable infrastructure

[2] We will, unless otherwise indicated, use the terms "genomic annotations" and "genomic tracks" interchangeably in this text, referring to the same group of data files, whether in context of visualisation, non–visual analysis, or elsewhere.

that improves the FAIRness of genomic annotations, with a particular focus on improving the end-user experience.

2. Value Proposition

Benefit to researchers: The FAIRification of Genomic Annotations WG aims to help researchers discover and access genomic annotations and integrate them from disparate repositories. Because annotations are often summarisations of primary data, the discovery of patterns and relations between annotation data is a cost-effective approach for data-driven research that can pave the way to novel discoveries based on subsequent and thorough analysis of the full datasets.

Lack of standardised metadata: Unfortunately, broader adoption of genomic annotations/tracks for exploratory analysis has been hampered by the lack of standardised metadata that – in line with the FAIR Principles – would enable the discovery and reuse of datasets.

Data portals: Important efforts to directly distribute tracks include data portals from larger consortia, such as ENCODE, and Functional Annotation of Animal Genomes (FAANG), mostly contributing to open science with well-annotated metadata. However, the metadata are provided according to distinct models and APIs, with varying levels of breadth and granularity.

Track hubs: Annotations generated in smaller research projects are often available in track hubs, listed on the UCSC Public Track Hubs page and the EMBL-EBI Track Hub Registry. However, these are mostly indexed at the track collection (hub) level only and the metadata of interest to research tends to reside on the level of individual experiments, samples, or data files. Furthermore, metadata for track hubs typically consists of one HTML webpage per hub, designed to be read by humans and as such lacks the consistent structure and semantics required for machine operability.

Genome annotations: A particular type of annotation file has, for historical reasons, been distributed independently of the rest, namely, genome annotations representing the initial characterisations of the individual regions of assembled genomes. Genome annotation data files (typically GFF files) denote coding and regulatory regions of DNA. The process of annotating an assembled genome can consist of computational predictions created by particular workflows/tools such as BRAKER, based on existing knowledge and/or related datasets, or provided by domain experts through a manual curation process supported by tools like Apollo. Since the annotation files are typically considered secondary data, they are often deposited alongside the genome assemblies. In many cases, however, lack of support from repositories has led researchers to publish the annotations elsewhere, such as in general repositories connected to journals. In both cases, there is typically limited metadata that relates to the annotation itself as an independent FAIR object, a problem acknowledged in a recent report by the Earth Biogenome Project, EBP.

Conclusions: All types of deposition and distribution of genomic annotations have particular challenges related to the existence of structured metadata, as well as heterogeneity in data structures and APIs. Hence, the mobilisation of such data by researchers in their specific analytical context is difficult due to a lack of standardised solutions.

Proposed impacts of the WG

Help researchers discover and integrate data: Allowing researchers to more easily leverage public data sources with genomic annotations through data-driven approaches, including several more specific use cases, such as:

- Importing annotation data into “genome browser” software for visual inspection or into other specialised and often non-visual analysis tools
- Generating data collections for training AI/ML models
- Improved discovery and accessibility of genome annotations from biodiversity projects, as well as compatible genome browser instances. This includes improved metadata connected to GFF files as FAIR objects, such as provenance information regarding the annotation workflows.

Start building a new public infrastructure: We will recommend an infrastructure built around persistently deposited metadata for genomic annotations in public repositories. A minimal metadata schema with strict requirements will ensure that the deposited metadata is consistently formatted in terms of structure and ontology values. Furthermore, recommendations related to maintainable metadata transformation pipelines, globally unique and persistent identifiers (PIDs), a metadata deposition registry, and a unified search API will define a relatively simple, scalable and maintainable public infrastructure that can then be expanded upon.

Improve existing infrastructures: Providing better integration of metadata into the existing infrastructure built around track hubs, which currently includes the UCSC Public Track Hubs page and the EMBL-EBI Track Hub Registry. The FGA-WG will develop a community perspective on how to improve the Track Hub Registry and the Ensembl genome browser services provided by EMBL's European Bioinformatics Institute (EMBL-EBI).

Community building and adoption: Initiating a general community around genomic annotation data and metadata, spanning biomedical and biodiversity subfields and beyond. Setting the stage to later include publishers and other FAIR initiatives to facilitate the adoption of the schema and infrastructure for published datasets, data sources, and software tools used for research findings.

3. Engagement With Existing Work in the Area

Previous and current efforts: Numerous efforts have been made to integrate data/metadata from multiple data portals and other sources into collections of genomic annotations, often in the context of analysis frameworks and tools (e.g. Cistrome, The Genomic HyperBrowser, LOLA, and HiCognition), genome browsers (e.g. the UCSC Genome Browser, Ensembl, JBrowse, and WashU Epigenome Browser) or domain-oriented databases (e.g. ChIP-Atlas, GWAS Catalog, RNAcentral, miRBase, Eukaryotic Promoter Database, ReMap, JASPAR, UniBind, Biocyc, and RegulonDB). A few (meta)data integration efforts have a more general scope and are of particular relevance to FAIRification of Genomic Annotations WG, including DeepBlue, The International Human Epigenome Consortium (IHEC) data portal, PanCancer Analysis of Whole Genomes (PCAWG), Human Cell Atlas, GenoSurf by the Data-Driven Genomic Computing (GeCo) project, the Mass Genome Annotation (MGA) repository, the AnnotationHub in the BioConductor project, the FAIRtracks project, PEPhub, and The Alliance of Genome Resources. We intend to build on existing efforts. Working group members have also previously carried out landscape reviews and harmonisation efforts that will contribute to this process[3].

Build on other Open Research initiatives: In addition to these domain-centric efforts, we aim to engage with other related initiatives within RDA and in other contexts, including:

Biodiversity: The outputs of the FGA-WG should be relevant for the Biodiversity Data Integration IG

CARE principles for Indigenous Data Governance: Endorsed and maintained by the Global Indigenous Data Alliance (GIDA), the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) are process-oriented principles aimed to complement the more data-oriented FAIR principles. The aim of the CARE principles is to balance the emphasis on greater data sharing with the needs of Indigenous Peoples to assert greater control over the application and use of indigenous data and knowledge for collective benefit. We will investigate how the CARE principles can be applied to the FGA-WG outputs, in coordination with the International Indigenous Data Sovereignty IG. In particular, fields in the harmonised metadata model that describe limitations on data usage will be aligned with the CARE principles. We will also investigate the use of the Labels and Notices from the Local Contexts project.

Data crediting: Data packages should be associated with credit metadata that retains data ownership and contributions to the data product, as well as any PIDs associated with data generation (funder ID, ORCIDs, etc). Relevant solutions include credit metadata schema and Apicuron

Data granularity: As genomic annotations are typically single files contained within datasets, our infrastructure is relevant as an example of granular data discovery at the sub-dataset level. We thus aim to investigate the adoption of the outputs of the Data Granularity WG.

---

[3] E.g. Bernasconi, Anna, et al. "The road towards data integration in human genomics: players, steps and interactions." *Briefings in Bioinformatics* 22.1 (2021): 30-44, Gundersen, Sveinung, et al. "Recommendations for the FAIRification of genomic track metadata." F1000Research 10 (2021), Sheffield, Nathan C., Nathan J. LeRoy, and Oleksandr Khoroshevskyi. "Challenges to sharing sample metadata in computational genomics." Frontiers in Genetics 14 (2023): 1154198.

ELSI aspects: The EOSC-Future/RDA Artificial Intelligence and Data Visitation Working Group (AIDV-WG) deliverables can contribute to supporting the ELSI aspects of this project.

FAIR data maturity: We will investigate the adoption of the recommendations from the FAIR Data Maturity Model WG to assess the implementation level of the FAIR data principles for the outputs of the FGA-WG.

FAIR Mappings: We will follow the initiative for a new RDA WG on FAIR metadata mappings. Metadata mappings/crosswalks will be a central aspect of our deliverable *2.1: Guidelines for enabling scalable and maintainable metadata transformation pipelines*, which in our case will be limited in scope to metadata about genomic annotations. We intend to build a fruitful collaboration with the FAIR Mappings WG, possibly representing a more data-centric and bottom-up approach to automated metadata mapping through e.g. the Omnipy Python library. We also aim to investigate the adoption of the outputs of the Brokering Framework WG in this context, as well as SSSOM and other initiatives related to metadata mapping.

Metadata and vocabularies: Relevant RDA groups include the Metadata IG and the Vocabulary Services IG.

Persistent identifiers and FAIR Digital objects: Relevant RDA groups include the PID IG, the FAIR Digital Object Fabric IG, and possibly the Data Versioning IG. Other relevant initiatives include the Decentralized PID Working Group from DeSci Labs.

Research evaluation aspects: Cooperation with the CoARA Working Group on 'Ethics and Research Integrity Policy in Responsible Research Assessment for Data and Artificial Intelligence'. Creating reliable and indexable metadata about genomic annotations will contribute directly to the aims of improving research evaluation for those working in the life sciences.

WorldFAIR: the CODATA-led 'Global Cooperation on FAIR data policy and practice (WorldFAIR)' project contributes to the implementation of the FAIR principles across domains, including in the life sciences. RDA also contributes to this project.

Other domains: While the Harmonised terminologies and schemas for FAIR data in materials science and related domains WG is focused on a different domain, the FGA-WG will be operating in parallel to ours and we believe there is a potential for some interesting overlaps between the two.

Interaction with GA4GH: Furthermore, we aim to follow, contribute to and investigate the adoption of relevant products by the Global Alliance for Genomics & Health (GA4GH), in particular refget/Sequence Collections, the Experiments Metadata Standard and Variant Annotation, all based on existing engagements by members of the FGA-WG. Other relevant GA4GH products include the Sequence Annotation ontology and the Data Use Ontology, both of which have already been adopted in the current recommendations from the FAIRtracks project.

4. United Nations Sustainable Development Goals (SDGs)

Zero hunger (SDG 2): The FGA-WG will contribute to this goal as it is related to plant biology. Target 2.5 relates specifically to the genetic diversity of seeds, cultivated plants and farmed/domesticated animals; research on these topics will all benefit from FAIR metadata on genomic annotations.

Good health and well-being (SDG 3): The FGA-WG will contribute to this goal as it is related to several biomedical use cases, including neonatal mortality (Targets 3.2 and 3.4), epidemics/diseases (Target 3.3), and death caused by various expositions (Target 3.9), where employing a large amount of FAIR metadata on genomic annotations is important.

Quality Education (SDG 4): The FGA-WG's emphasis on standardisation and FAIRification will contribute to providing more just and equal opportunities for all to acquire information, thereby reducing inequalities concerning education and learning.

Industry, innovation and infrastructure (SDG 9): The output of the FGA-WG will facilitate research improving accessibility and usage of public data. This might in particular benefit researchers from low-income countries, as reuse of public data should reduce the need to produce new data (see Target 9.5).

Life below water (SDG 14) and **Life on land (SDG 15)** are relevant for all the applications where FAIR metadata on genomic annotations can be used in biodiversity contexts.

5. Work Plan

Main challenges: The two main challenges the FAIRification of Genomic Annotations WG will tackle are as follows:

1. Harmonising metadata about genomic annotations and their datasets through a *minimal metadata schema*; and

2. Developing a *maintainable and scalable infrastructure* that makes use of the schema to provide solutions for metadata transformation, validation, sharing and discovery

Continuation of previous efforts: The FAIRtracks ecosystem is an existing infrastructure that begins to address both of these aims, built in the context of ELIXIR Europe (FAIRtracks.net website / F1000Research blog on FAIRtracks). Building on this infrastructure as well as other relevant solutions, such as IHEC and GenoSurf, allows the ambitious goals of the FGA-WG to be feasible within the planned 18-month period. An important first step is to harmonise the FAIRtracks schema with relevant data models and solutions, as well as improve the harmonised schema according to the requirements of the main use cases of the FGA-WG. This will build on expertise within the FGA-WG itself, but also adopt relevant standards and recommendations from other Working Groups in the RDA, GA4GH, and other initiatives (see also Section 3).

5.1. The final Recommendation of the WG

Deliverable 1. Recommendations for a Minimal FAIR Metadata Schema

A minimal FAIR metadata schema for genomic annotations that builds on the existing FAIRtracks metadata schema, further harmonised with relevant data models such as from the IHEC, GeCo and FAANG projects, and evolved to support the FGA-WG use cases. The deliverable will consist of a document outlining general recommendations for supporting the FAIRness of metadata on genomic annotations. This will include an overview of the structure and design of the recommended metadata schema. The schema itself will be made available as a release of a version-controlled software repository.

Deliverable 2. Core Infrastructure Recommendations

Recommendations on how to adopt the metadata schema (Recommendation 1) to build an infrastructure to improve the FAIR-ness of genomic annotations in practice. We will evolve the recommendations in parallel to building an actual infrastructure that adopts them, following the proven principles expressed in the Manifesto for Agile Software Development. The core infrastructure recommendations should include:

> *2.1. Guidelines for enabling scalable and maintainable metadata transformation pipelines:* To allow for the continuous evolution of e.g. metadata schemas, ontologies, or data sources, there is a need to define best practices for scalable and maintainable pipelines that transform metadata from existing sources to fit our minimal schema. The new Omnipy Python library is a possible framework for developing and orchestrating

such metadata mapping/transformation flows; it has been designed with this exact use case in mind. We will also build on experience and code from other projects, such as the ones listed under Recommendation 1 above. We furthermore aim to contribute to and exchange experiences with the upcoming FAIR Mappings WG.

2.2. Strategy for persistent and public deposition of harmonised metadata: Community trust in the availability of FAIR metadata on genomic annotations depends on a common strategy for permanent storage of such metadata in public repositories. We will also need to define and implement a simple registry of relevant metadata depositions.

2.3. Strategy for providing persistent identifiers: Adopting the FAIR principles necessitates persistent identifiers (PIDs) on at least the collection level, preferably on a file level. As a minimal solution, we can adopt the current FAIRtracks solution of a combination of a DOI for a metadata deposition and a locally unique identifier within that dataset. We will investigate the option of producing content-derived digests for this purpose, building on experience from the Sequence Collections GA4GH working group and elsewhere. Other alternatives include adopting decentralised PIDs from the dPID Working Group from DeSci Labs.

2.4. Define uniform search API for downstream services and end users: To facilitate adoption by end users and integration with software tools, we will define a standardised API for downstream search and discovery that works independently of any particular implementation of a search service. This will build on experience from the GenoSurf, TrackFind, IHEC data portal and similar search services as well as from downstream users of such APIs, such as developers of analysis software (e.g. LOLA and GSuite HyperBrowser), as well as direct use by analytical end users.

Deliverable 3. Proof-of-concept integrations of third-party services with core infrastructure

These integrations will be primarily developed through the initiative of FGA-WG members and in their particular research context as integrations with other projects. The work that should be counted towards the FGA-WG entails mainly coordination and exchange of ideas and experience, as well as the improvements and modifications needed on the infrastructure side to facilitate the integrations. The deliverable will be in the form of a report.

3.1. Metadata transformation pipelines: Data pipelines (following Recommendation 2.1) developed to transform metadata from particular data sources to conform to the minimal metadata schema (Recommendation 1) and deploy compliant metadata which is made uniquely identifiable with PIDs (Recommendation 2.3) into permanent public storage (Recommendation 2.2).

3.2. Search service implementations: At least one search service that imports metadata from permanent storage (Recommendation 2.2) and implements the standardised search API (Recommendation 2.4).

3.3. Tool integrations: Integration of the search API in downstream tools and libraries.

3.4. Track hub integrations: Improved integration with the track hub-based infrastructure to improve the generation, discovery and distribution of data hosted as track-hubs. This

will build on the existing collaboration with the Track Hub registry and the Ensembl genome browser through the FAIRtracks project, but will also be open for other collaborations.

Deliverable 4. Resources that support community-building and adoption

Resources that support building and maintaining a community and facilitate the adoption of the recommendations and related infrastructure (see sections 5.5 and 6 for more details). The deliverable will be in the form of a report.

5.2. Milestones, code or other deliverables that will be developed

Gantt chart: Please refer to the Gantt chart in Appendix 1 for an overview of the FGA-WG's deliverables and tasks, and their schedule. A first milestone, consisting of the first version of the harmonised metadata schema, is planned to be completed within the first six months of operation. The second milestone, consisting of the initial draft of the core infrastructure recommendations, is planned to be completed within fifteen months.

Publication of outputs: Deliverables will be published in public repositories, with assigned DOIs. Code will be provided through public version control systems, with releases made available in alignment with the recommendations by the FAIR for Research Software (FAIR4RS) WG. The produced artefacts will, in particular, be shared on FAIRsharing as recommended by the FAIRsharing Registry: Connecting data policies, standards and databases RDA WG. We will also investigate sharing the outputs in other registries recommended by the Life Science Infrastructures, the European Open Science Cloud (EOSC) and other entities promoting FAIR practices for research outputs.

5.3. The WG's mode and frequency of operation

Regular monthly meetings: The working group will meet virtually every month to present members with updates on tasks and from co-chairs and facilitate open work sessions. The meetings will be 60-90 mins in length, chaired by one of FGA-WG co-chairs with organisational support from RDA TIGER (including but not limited to agenda setting, monitoring actions, upcoming events of interest, outreach to related projects, etc.). The regular monthly meetings will target a time within typical working hours of our diverse member time zones, such as 8 am Pacific / 11 am Eastern / 4 pm UTC, to maximise the number of people who find the time convenient.

Task-specific meetings: Additionally, smaller, individual meetings focused on specific tasks will be organised where relevant. These meetings may also be timed to accommodate more desirable time zones for subsets of the working group membership. Co-chairs/task leads will meet outside the regular monthly meetings where necessary, with coordination support from RDA TIGER.

5.4. Plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation

Member contributions: Through the regular monthly meetings and task-specific meetings, FGA-WG members will have the opportunity to provide their input in these 'face-to-face' virtual settings, offline via feedback/comments on shared documents or the FGA-WG mailing list. All members will be invited and encouraged to contribute to the tasks and activities of the working group.

Consensus-based: To the extent practically possible, we aim to reach a consensus on the various decisions that will make up the recommendations. If there are differing opinions on FGA-WG's aims, tasks or methods of work, the co-chairs will decide on the most appropriate course of action, providing documentation and an explanation of their decision to all FGA-WG members. The FGA-WG will uphold the RDA Code of Conduct.

Task leadership: Co-chairs will take the lead for Deliverables and their dependent Tasks (see Gantt chart), providing updates to the FGA-WG members at the regular meetings.

5.5. Planned approach to broader community engagement and participation

The FAIRification of Genomic Annotations WG aims to foster extensive community engagement by actively involving various stakeholders. The following is a list of potential actions and activities for the FGA-WG to engage with the broader community:

Invite broadly to attract members: Ideally, we would like to have members with a wide variety of roles, such as data producers, domain experts, tool/service developers, FAIR/RDM specialists, ethics and ELSI expertise, analytical end users, and anyone interested that do not fall into any of these categories. We will of course target existing RDA members, but our main efforts will be spent on inviting individuals and groups in the life science community in general, regardless of previous association/knowledge of RDA. In this way, we believe the FGA-WG could attract several new members to RDA, some of whom would probably also want to engage with other groups and activities.

Presentations to promote alignment: We aim to invite representatives from key projects and organisations to participate in our meetings and presentations, as well as suggest presentations of our activities and outputs in their fora. Whenever mutual interest with external groups has been established, then we intend to actively collaborate with these entities to ensure alignment and maximise impact.

Particular plans per use case: The three use cases necessitate engagement with different types of external entities.

Biomedical analysis community and data providers: For the first, biomedically oriented use case, we will contact A) research groups and other developers of tools and methodologies that would benefit from integrating with the outputs of the FGA-WG, and B) current data-producing projects,

as well as maintainers of outputs from finalised projects, such as data portals. As a motivation for both types of entities, we believe the contribution to/integration with the FGA-WG outputs might lead to increased usability and usage of their tools/datasets.

Biodiversity community: For the second, biodiversity-oriented use case, we will primarily contact biodiversity projects, as well as developers of tools and frameworks used in this context. As most large biodiversity projects are affiliated with the Earth Biogenome Project (EBP), we will in particular strengthen the existing contact with the EBP Annotation Subcommittee, as well as with individual projects and communities such as the European Reference Genome Atlas (ERGA), EBP-Nor, and the ELIXIR Biodiversity community. We have not yet enlisted a co-chair with particular responsibility for the biodiversity use case, so there is a need to emphasise engagement with this community early on. While it would be beneficial to add support for needs specific to the biodiversity use case in the initial definition of the metadata model, our work plan allows for iterative extensions of the model throughout the project period (see Gantt chart in Appendix 1).

Genome browser teams and Track Hub providers: The third use case is dependent on collaboration with providers and developers of genome browsers and Track Hub-related services. Through the FAIRtracks project, we have already established a fruitful collaboration with the Genome Analysis team at EMBL-EBI, whose services include the Track Hub Registry and the Ensembl genome browser. We aim to also get in touch with other genome browser developers and providers, importantly the UCSC Genome Browser team, as well as other groups like the Alliance of Genome Resources and the JBrowse team. We are also interested in getting in touch with projects or individuals producing Track Hubs.

Anchor the FGA-WG within Life Science infrastructures and other projects: Of particular importance are the Life Science infrastructures currently represented in the Life Science Data Infrastructures IG: ELIXIR, The Australian BioCommons, NIH Office of Data Science Strategy, and H3ABioNet. The FGA-WG is currently coordinated with ELIXIR through the existing FAIRtracks ecosystem, which is part of the contractually bound Service Delivery Plans (SDP) from both ELIXIR Norway and ELIXIR Spain towards life science researchers in Europe. FAIRtracks has also been awarded the quality mark "ELIXIR Recommended Interoperability Resource". Plans will be made on how to merge the FAIRtracks project with the FGA-WG in a way that maintains the strengths of the existing structures while allowing global collaboration to take form, possibly anchored also within the other large infrastructures. Similar structural considerations will be investigated for other projects that are interested in joining forces.

Make use of RDA plenaries and other events: We will make use of RDA plenaries to engage with other RDA groups as indicated in section 3, to share experiences and investigate the adoption of relevant RDA recommendations. Other relevant events include the ELIXIR Biohackathon and other workshop or hackathon-type events.

Coordinate with GA4GH: We will, as detailed in section 3, engage with related efforts within the GA4GH, present our outputs in GA4GH meetings, and make use of our connections within that organisation to attract new members to the FGA-WG. In general, adding new contact points between these two organisations might help increase adoption across the organisations and help reduce parallel development efforts.

Operate community resources: As we intend to build a vibrant community dedicated to enhancing the FAIRness of genomic annotations, we aim to implement and host resources that

support building and maintaining a community, such as web pages, documentation, chat, mailing lists, and social media accounts. One option is to build on the efforts already made to implement the FAIRtracks.net website, transforming this into a community portal/hub for the FGA-WG.

Visibility in conferences and journals: The community could be further expanded through the organisation of special sessions at Bioinformatics-related conferences (such as SWAT4HCLS, ISCB-ISMB, IEEE BIBM, IEEE BIBE, IWWBIO, etc.) and of special issues in bioinformatics journals.

Open strategy meetings: While most meetings will be working and/or reporting meetings, there is a need for more strategy-oriented meetings, at least for the co-chairs. One possibility is to, once in a while, expand these into open strategy meetings. The focus of such meetings could be to coordinate engagement with external groups, investigate extra funding sources, coordinate grant applications, organise workshops, facilitate participation in conferences, etc. Such meetings could be a way to engage members, PIs, external groups and others who are interested in contributing to the FGA-WG but are not able to follow the day-to-day developments.

Other relevant groups and projects: The Galaxy community, AgBioData, International Society for Biocuration, European Molecular Biology Organization (EMBO), and many more.

6. Adoption Plan

This Adoption Plan outlines a streamlined approach for implementing the Recommendations of the FAIRification of Genomic Annotations WG and fostering widespread adoption within member organisations and the broader community. The Adoption Plan aims to accelerate the adoption of FAIR data practices for genomic annotations by providing support, training, and collaboration opportunities to stakeholders. Through targeted efforts, the FGA-WG aims to drive positive change in data management practices and promote a culture of data sharing and accessibility.

Internal Implementation: FGA-WG members will lead the integration of the recommended metadata schema, ontologies, and metadata transformation pipelines within their organisations. This includes facilitating data deposition to public repositories and implementing Persistent Identifiers (PIDs) for data traceability.

Training and Capacity Building: Capacity-building initiatives will empower stakeholders with the skills needed to make effective use of the core infrastructure. Workshops and webinars will provide practical guidance, with a focus on collaboration with platforms like Galaxy and BioConductor.

Engagement with Communities: Outreach efforts will target relevant communities and initiatives, leveraging domain organisations such as the Life Science Infrastructures, EBP, EMBL-EBI, BGI, CNGB, Genomics England, Genomics Australia, GA4GH, and EMBO in addition to the RDA (see section 5.5). Participation in conferences and forums will amplify awareness and uptake of the FGA-WG's recommendations.

Documentation and Communication: Comprehensive documentation and user-friendly resources will support adoption efforts. Communication channels will be utilised to disseminate information and provide ongoing support to stakeholders.

Feedback and Iterative Improvement: Feedback mechanisms will drive iterative updates to the data model and the core infrastructure, ensuring alignment with evolving community needs.

External adoption: We aim to enable downstream adoption via third-party grants provided by the RDA TIGER project toward the end/after completion of the FGA-WG. In particular, we aim to approach journals and repositories to propose adoption, something which might go a long way towards turning the outputs into a de facto standard.

Further development into an endorsed standard: A successful project outcome will constitute significant progress toward a standard for genome annotation metadata. One possibility for developing the recommendations into an endorsed standard is to apply for it to become a GA4GH product. We will also, with the help of RDA TIGER, investigate other paths for further development into an open science standard.

7. Initial Membership

Co-chairs

Name	Institution	Email
Sveinung Gundersen	ELIXIR Norway / University of Oslo	sveinugu@ifi.uio.no
Anna Bernasconi	Politecnico di Milano	anna.bernasconi@polimi.it
Nathan Sheffield	University of Virginia	nsheffield@virginia.edu
Adam Wright	Ontario Institute for Cancer Research	adam.wright@oicr.on.ca

Members

Name	Institution	Email
~~Peter Harrison¶~~	~~EMBL-EBI¶~~	~~peter@ebi.ac.uk¶~~
Beth Flint	EMBL-EBI	beth@ebi.ac.uk
David Bujold	McGill	david.bujold@mcgill.ca
Katharina Hoff	University of Greifswald	Katharina.hoff@uni-greifswald.de
Ole K. Tørresen	EBP-Nor / University of Oslo	o.k.torresen@ibv.uio.no
Philipp Bucher	SIB (Switzerland)	philipp.bucher@sib.swiss
Neko He	China National Genebank (CNGb), Shenzhen, China	heziyi@genomics.cn
Nicholas Owen	University College London (UCL)	n.owen@ucl.ac.uk
Anthony Bretaudeau	INRAE, France	anthony.bretaudeau@inrae.fr
Francis P. Crawley	GCPA & SIDCER, Leuven, Belgium	fpc@gcpalliance.org
Timothee Cezard	EMBL-EBI	tcezard@ebi.ac.uk
Federico Bianchini	ELIXIR Norway / University of Oslo	fredebi@uio.no

Name	Institution	Email
Pável Vázquez	ELIXIR Norway / University of Oslo	p.v.faci@medisin.uio.no
Nazeefa Fatima	ELIXIR Norway / University of Oslo	nazeefaf@uio.no
Jeanne Chèneby	ELIXIR Norway / University of Oslo	jeannech@uio.no
David Molik	USDA ARS ABADRU	david.molik@usda.gov

Acknowledgements

The proposed FAIRification of Genomic Annotations WG was pre-selected as a pilot to receive support by the Research Data Alliance facilitation of Targeted International working Groups for EOSC-related Research solutions (RDA TIGER) project, which started January 2023. We are very grateful for the support we have received so far from Ryan O'Connor, Alexandra Delipalta, Ari Asmi, and others.

A special thanks to Peter Maccallum, CTO of ELIXIR, for connecting this initiative to the RDA TIGER application, and to Jeff Christiansen, Deputy Director of Australian BioCommons, for facilitating meetings "down under".

Appendix 1: Overview of the FGA-WG's deliverables and tasks, and their schedule