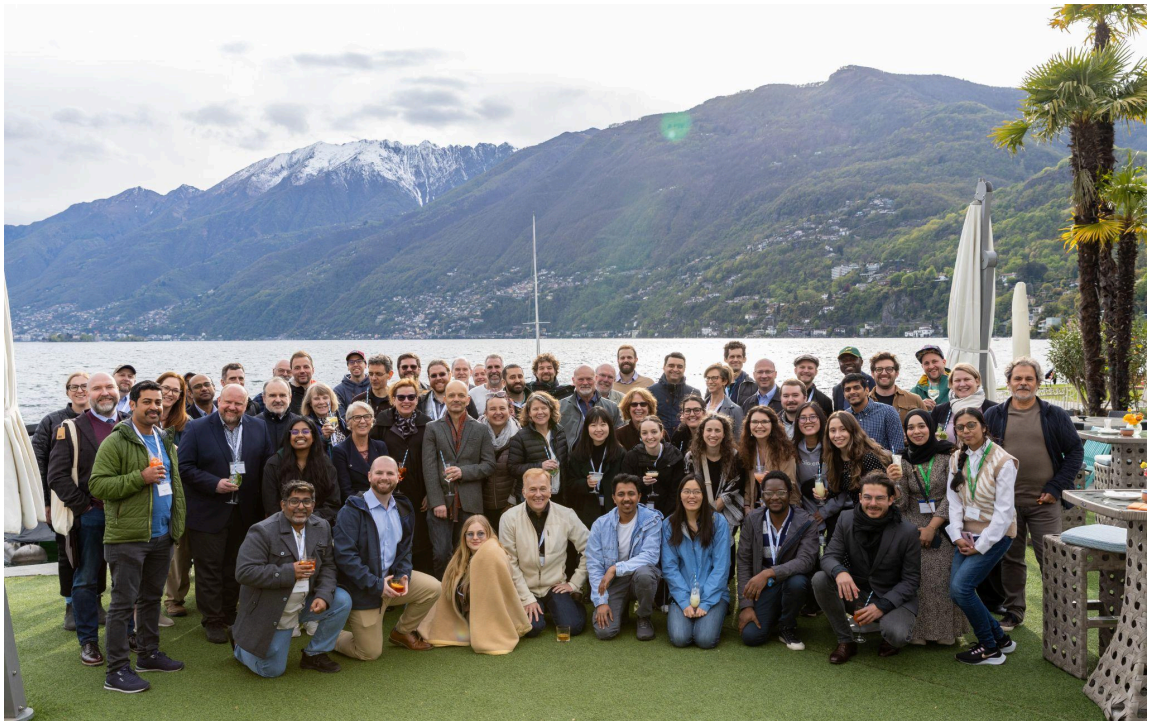
2024 GA4GH April Connect meeting report



The Global Alliance for Genomics and Health (GA4GH) gathered for the 2024 April Connect meeting in Ascona, Switzerland and online from 21 to 24 April. The GA4GH Connect meetings provide an opportunity for contributors to advance the GA4GH Road Map, showcase GA4GH standards and policies in action, and gather feedback on product development and community needs. The meeting brought together 103 in-person attendees and 312 virtual attendees for updates from Work Streams and Driver Projects, breakout sessions, and themed events.

Table of contents

Monday, 22 September

- Plenary session
- Ethical Provenance workshop
- Phenopackets: horizons and perspectives
- GKS: releases are forever
- Driver Project workshop: AI / ML
- GA4GH Testbed Infrastructure
- GKS: casino Cat-VRS
- Experiments Metadata Standard
- Driver Project workshop: data discovery
- Command line interface for GA4GH environments
- Neuroscience Community of Interest meeting
- Passports in production
- Rare Diseases Community of Interest meeting

Tuesday, 23 September

- Unveiling GA4GH's five-year Equity, Diversity, and Inclusion Strategic Plan
- Around the world in one query
- DRS v1.5 key features and plans for 2024
- Beacon thematic working session: cohorts
- Technical Assignment Subcommittee (TASC) meeting
- Data visiting
- Beacon variants
- GKS: the feature is not enough

Wednesday, 24 September

- Implementation of Beacon in cancer use cases
- GKS: you only implement twice
- Beacon resolved: a session for development
- AAI & Passports "what's next?" unconference sessions
- Crypt4GH, workflows, and key management
- Federated cohort building
- GKS: from the roadmap with love
- Beacon filter solutions
- DaMaSC Schema Registry
- Driver Project workshop: multi-site geography collaboration and governance

---

Sunday, 21 April 2024

VCF v4.5: scalability and methylation

Agenda and slides • Recording

Attendees discussed new features, fixes, and clarifications in the VCF v4.5 draft specification. These include local alleles designed to improve VCF scalability, draft methylation tags, and the suite of test VCF files that are now part of the hts-specs repository.

Key takeaways

- Attendees shared live feedback on the VCF v4.5 draft specification.

Next steps

- Review and incorporate feedback into a new version in June, upon the completion of a six week public feedback phase.

Regulatory & Ethics Work Stream (REWS) general meeting

Agenda and slides • Recording

Attendees heard updates on REWS projects and initiatives. The session also featured a presentation on the *Draft WHO principles for human genome data access, use, and sharing*.

Key takeaways

- Study Groups and product teams are at the heart of REWS. The Work Stream has finalised a new approval process that all REWS products will adhere to.
- REWS has submitted a response to the *Draft WHO principles for human genome data access, use, and sharing.*

Next steps

- Share the new REWS road map with the GA4GH community. This is a great opportunity for contributors to join new REWS initiatives.

GA4GH Implementation Forum (GIF)

Agenda and slides • Recording

Attendees learned about and provided input on the new GA4GH Implementation Forum (GIF). GIF aims to facilitate standards implementation, establish interoperability, and promote best practices to ensure that GA4GH solutions can solve real-world problems. The group discussed ways that GIF can address barriers to the implementation and adoption of GA4GH products.

Key takeaways

- GIF has three core tenets: implementation, interoperability, and inclusivity.
- GIF is in the process of collecting feedback on its structure and proposed activities to ensure that it meets the needs of the community.
- Proposed activities include hosting showcases and hackathons, facilitating projects, and developing implementation resources.

Next steps

- Gather feedback on the direction of GIF. The GA4GH community is invited to add comments on the Miro Board.
- Circulate the draft GIF Charter to ensure GIF can help address the needs of the community.

- Follow-up with and assess current status of GIF projects, updating the GIF webpage as needed.

Driver Project workshop: data harmonisation

Agenda and slides • Recording

This workshop was designed to kickstart important discussions, encourage collaboration, and pinpoint common challenges and themes encountered in Data Harmonisation. Attendees, with a focus on Driver Project members, had the chance to exchange insights and best practices and explore opportunities for collaboration. The session was highly participatory, featuring breakout discussions for deeper exploration of specific topics.

Key takeaways

- Workshop leads Mónica Muñoz Torres, Heather Creasy, and Brian O'Connor engaged Driver Projects and other attendees in discussions around tool portability, metadata harmonisation, and data and evidence provenance.
- Discussions produced the following topics for further exploration:
    - translating harmonisation tools between different workflow languages;
    - standardising how to describe workflow inputs and outputs for Cloud APIs;
    - determining how to bring in standards created for the Rare Disease space by the GREGoR consortium into GA4GH;
    - understanding the work needed to transfer metadata into a harmonised model;
    - standardising how we track genomic interpretations and how they are used in downstream and reclassification tasks.
- Some topics discussed overlapped with existing GA4GH and external groups, such as the Data Model and Schema Consensus (DaMaSC) group, the RDA FAIR Data Maturity Model Working Group, and the Experiments Metadata Standard group. Coordination and alignment will occur to avoid duplicating efforts.

Next steps

- Develop an action plan to move forward with identified topics of interest.

Opportunities and obligations in conducting responsible genomics research: role-based perspectives

Agenda and slides • Recording

Attendees discussed obligations and constraints faced by those with different roles in the genomics research ecosystem and identified points for a future manuscript.

Key takeaways

- Researchers, funders, ethics boards and other oversight bodies, journal editors, and community partners have crucial roles in the genomic research ecosystem.
- This project, along with developing a manuscript, can help identify the ways in which GA4GH can facilitate an effective genomic research ecosystem.

Next steps

- Conduct a landscape analysis, finalise a methodology, and define the scope for a manuscript.

Federated variant level matching

Agenda and slides • Recording

Attendees explored current challenges and opportunities in Beacon variant matching, discussed strategies to prepare data for variant matching, examined phenotype sharing policies, and interrogated the implications of public vs registered access.

Key takeaways

- Session participants explored options, benefits, and drawbacks of different strategies in building variant stores to support federated variant-level matching.

Next steps

- Develop a best practices guide for implementing variant-level matching.

Data Model and Schema Consensus (DaMaSC) style guide and ontologies

Agenda and slides • Recording

Attendees discussed and defined a technical style guide to enhance consistency, clarity, and interoperability across GA4GH. The aim was to share best practices for expressing data model schemas and bridge vocabulary and representation across different GA4GH product teams and standards. Attendees came to an agreement on the definition and purpose of the style guide, reviewed use cases, and discussed the scope of this project.

Key takeaways

- The group agreed on several needs, including:
    - standardising population descriptors and the importance of aligning terminology with scientific understanding and social sensitivity;
    - collaborating and aligning across various GA4GH product teams to establish best practices for capturing data provenance and lineage, and ensuring consistency and interoperability across the GA4GH community.
- Examining real-world use cases can ensure the usefulness of any recommendations or best practices.

Next steps

- Collaborate with the Technical Alignment Subcommittee (TASC) on the project itself and potential implementation of any best practices.
- Develop a best practices guide with next steps to define the scope and recommendations.

Monday, 22 April 2024

Opening session

Slides • Recording

The opening session featured an introduction to the meeting and welcoming remarks from Michael Baudis and Christophe Dessimoz about ongoing data sharing work in Switzerland. Heidi Rehm, Chair of GA4GH, presented an introduction to GA4GH, providing an overview of the key work of the organisation. This was followed by presentations from new Driver Projects, including Biodata Catalyst (BDC), Genomic Data Infrastructure (GDI), NIH Cloud Platform Interoperability (NCPI) effort, International Precision Child Health Partnership (IPCHiP), and the Biomedical Research Hub (BRH).

Ethical Provenance Workshop

Agenda and slides • Recording

Attendees reviewed the final list of Data Access Agreement clauses resulting from the culmination of a yearlong empirical project funded by the US National Institutes of Health (NIH).

Key takeaways

- Clauses related to IP, liability, scientific publication, and contract modification should be understandable and transparent to increase the harmonisation of data access agreements amongst researchers, research communities, and institutions.

- It is still important to balance the need to protect data providers' interests and downstream users' rights to ensure more widespread access to data.

Next steps

- Revise data access agreement clauses and circulate to the GA4GH community for feedback.
- Continue work with Sage Bionetworks, a trusted nonprofit leader on data sharing and reuse, on public-facing explanatory notes and a guide to understand what the clauses enable from different perspectives.
- Update the Data Sharing Lexicon.

Phenopackets: horizons and perspectives

Agenda and slides • Recording

Attendees discussed the evolving landscape of Phenopackets and how to capitalise on the strengths of the standard, leveraging the presence of Driver Project representatives to identify opportunities, address challenges, and brainstorm potential solutions.

Key takeaways

- The group discussed the critical role of Phenopackets in facilitating the exchange of structured phenotypic data for computational analysis and interoperability.
- Various tools and implementations developed for Phenopackets, such as the Java library, Phenopacket Store, and Phenopacket Lab, were showcased. The Bento platform was also introduced, demonstrating data ingestion and retrieval using Phenopackets.
- Diverse use cases and implementations of Phenopackets were shared, from diagnosing undiagnosed diseases in Japan to managing data release processes in Australian Genomics.
- Attendees identified and discussed gaps and areas for improvement in the current specification, such as representing consent, cancer data, spatial data, and study-specific information.

Next steps

- Collate detailed feedback from participants on their experiences with Phenopackets to identify gaps, challenges, and evolving requirements. This information will help inform and prioritise the project's road map, ensuring that the standard meets the diverse needs of the community.
- Continue developing and refining tools for Phenopackets, such as the Phenopackets Java library, Phenopacket Store, and Phenopacket Lab.

- Increase outreach and educational efforts to promote the understanding and adoption of Phenopackets within various projects and organisations, emphasising the schema's flexibility and how it can be extended for specialised fields, such as cardiology or oncology.

GKS: releases are forever

Agenda and slides • Recording

Attendees discussed and reviewed updates to the GKS maturity model and release process for supporting transparent and scalable standards development. The group focused on how the process applies to developing the Variation Representation Specification (VRS), Variation Annotation (VA) specification, Sequence Annotation (SA) specification, and Categorical Variation Representation Specification (Cat-VRS).

Key takeaways

- The VA team shared an overview of the core information model and the process of profiling for specific use cases, such as variant pathogenicity.
- The VRS team has implemented a new approach, called reference length expression, to handle repeating elements more effectively. This method enables compact and numerically encoded representation of sequences with repeating elements, addressing previous issues with long state sequences.

Next steps

- The SA team will identify core elements, such as sequences, features, locations, and relations, to develop a core sequence annotation model.
- The Cat-VRS team will develop a terminology, data model, and referenceable notation for representing categorical variation. Its goals include enabling efficient matching between assays and categories and developing a flexible data model.
- Initiate a calendar-based version release schedule across all four GKS products to facilitate the promotion of certain features to trial use.
- Promote the implementation of two feature development processes:
    - proposing and discussing new feature ideas through GitHub's discussion board;
    - tracking progress using a GA4GH project board, showcasing upcoming, in-progress, and completed tasks.

Driver Project workshop: AI/ML

Agenda and slides • Recording

Attendees discussed opportunities and challenges for GA4GH in the area of genomic standards development for artificial intelligence (AI) and machine learning (ML) use cases.

Key takeaways

- The Cloud Work Stream presented on assessing and evaluating whether Cloud standards are AI ready.
- A primer on retrieval-augmented generation in large language models was shared, providing a potential topic to explore whether the community can use these technologies to facilitate data discovery and reduce the burden of data harmonisation.
- The European Genomic Data Infrastructure (GDI) project presented on federated learning in GDI, sharing their use case questions.

Next steps

- Schedule a follow-up session with interested individuals to identify potential leads and conduct a landscape analysis on current activities in the AI/ML space with GA4GH overlap. This exercise could lead to the creation of a new Study Group on AI/ML.

GA4GH Testbed Infrastructure

Agenda and slides • Recording

Attendees learned about the current status of the harmonised testbed infrastructure, followed by a demonstration of the process for submitting a test report to the testbed and an exhibition of the results of these reports in a user interface. Details on the tooling available to build and manage test suites using existing frameworks were also presented.

Key takeaways

- It would be beneficial for researchers, developers, and institutes currently using or planning to adopt GA4GH standards to ensure their work is compliant as per specifications.
- It is important to increase awareness, knowledge, and exposure of the GA4GH testbed infrastructure and its role in improving visibility of GA4GH implementations.

- Contributors will be encouraged to write new test suites and compatibility suites.
- A list of potential enhancements to the existing infrastructure based on feedback from workshop participants will feed into the GA4GH technical team development process.

Next steps

- Analyse and evaluate the feedback from the audience on potential enhancements.
- Align the GA4GH technical team's road map with the potential enhancements and features for progressing conformance testing of implementations for standards.
- Prepare the development instance of the testbed for presentation at GA4GH 12th Plenary in September 2024.

GKS: casino Cat-VRS

Agenda and slides • Recording

Attendees discussed the development and challenges of the Categorical Variation Representation Specification (Cat-VRS) and Variation Representation Specification (VRS) for structural variations (SVs), aiming to create computable and dynamic models for representing genomic variations. During this working session, attendees reviewed the Cat-VRS v1.0 draft schema, including progress towards handling priority categorical variation areas — canonical variants, protein sequence variants, CNVs, and fusions — and organising the Cat-VRS repository, documentation, and github workflow.

Key takeaways

- Part 1: Categorical Variation Representation Specification goals and challenges
    - The primary objectives of the Cat-VRS group is to create a computable and automatable search process for identifying and querying classes of genomic variations, with the aim of developing a terminology and data model, a JSON schema implementation, and a reference Python implementation.
    - Key challenges include the complexity of categorising variants and the hierarchical relationships between them, with design goals focused on precise representation, flexibility, error resistance, and computability.
- Part 2: Structural Variation Representation
    - The design goals for VRS for structural variations (SVs) are focused on creating a file format capable of fully describing a genome, including

phased variants and ordering of variants within duplicated regions. The model uses low-level building blocks, primarily adjacencies to represent SVs, and aims to avoid inconsistencies with VCF.
    - Challenges include accurately classifying complex genomic rearrangements and the need for a comprehensive model that can handle categorical labels and detailed descriptions of structural variations.

Next steps

- Develop a comprehensive test set of categorical variants. Feedback and contributions are encouraged from Driver Projects and implementers to ensure a diverse and representative set of categorical variation.
- Address temporality and versioning in variant classification. The team will explore methods to handle dynamic updates to variant classifications, based on evolving knowledge and annotations.

Experiments Metadata Standard

Agenda and slides • Recording

This working session was held to continue development of a list of core properties to characterise sequencing assay types and draft categories to regroup these assays. Attendees were encouraged to look at the Scope Statement, which describes what metadata this group covers and the way the checklist will be made available.

Key takeaways

- Discussions revolved around the experiments metadata core properties checklist. Contributor feedback was given on the following topics:
    - adding fields such as kit retail name and kit manufacturer — there is a team at the German Human Genome-Phenome Archive (GHGA) working on this list;
    - using NCI Thesaurus identifiers and its definitions;
    - considering the approach to ontology and controlled vocabulary;
    - using "ID" as opposed to "name" to be more restrictive;
    - using CURIE format and utilising identifier.org.

Next steps

- Encourage group members to complete columns I and J in the core properties checklist, which will help with mapping the current list to your organisation's approach.

    - If your field does not map directly, or could relate to more than one property, leave a comment with an explanation.
- Seek out domain expertise to fill out the core properties checklist for various types of assays.

Driver Project workshop: data discovery

Agenda and slides • Recording

Attendees heard an overview of the Discovery Work Stream and product teams. To encourage alignment and synergy, Driver Project Champions presented on their respective initiatives.

Key takeaways

- There was interest in a Schema Repository for depositing and retrieving schemas. However, there are questions about whether the focus should be on defining schemas or on registering and finding existing schemas used in different implementations, alongside the issue of infrastructure responsibility and ownership.
- There was discussion about integrating Phenopackets within Beacon queries, which highlighted the need for clarity on how to support Phenopackets within the Beacon framework and challenges associated with ontology mapping.

Next steps

- Distil Driver Project needs that were presented at the session, and, considering priority, bandwidth, and feasibility, identify opportunities for product improvement within the appropriate Discovery product team.

Command line interface for GA4GH environments

Agenda and slides • Recording

Attendees learned about ongoing work to integrate confidential computing with other GA4GH standards.

Key takeaways

- The project — which is part of the ELIXIR BioHackathon Cloud — has attracted a group of Google Summer of Code 2024 contributors.

Next steps

- Submit this project to the GA4GH Implementation Forum (GIF) to ensure compliance and alignment with other GA4GH standards.

Neuroscience Community of Interest meeting

Agenda and slides • Recording

Neuroscience Community Co-Chairs introduced the community structure and theme leaders. Theme leaders presented on the topics of scientific collaboration and education, data governance, and data harmonisation.

Key takeaways

- Neuroscience Community Theme Leads – Nick Halper, Kim Ray, and Jean-Baptiste Poline – presented on three distinct themes and related project ideas. Teams are encouraged to form around the themes of scientific collaboration and education, data governance, and data harmonisation.

Next steps

- Once formed, theme teams will meet regularly to identify and work on a community project aimed at furthering interoperability and collaboration.
- Teams will share project updates at an upcoming Neuroscience Community meeting, which will be hosted by either the Brain Research International Data Governance and Exchange (BRAIN)/International Neuroinformatics Coordinating Facility (INCF) or the Society of Neuroscience.
- Community members are encouraged to **sign up** for theme teams that they are interested in joining.
- Once members have signed up, theme leads will set up their initial theme team meeting. In these meetings, theme teams will continue discussions about the specific project they plan to pursue.

Passports in production

Agenda and slides • Recording

Attendees learned about current examples and uses of Passports in production within different environments and systems. Feedback was collected to help shape discussions for a future Driver Project workshop on data access authorisation.

Key takeaways

- The US National Institutes of Health (NIH) Researcher Auth Service (RAS) implementation of Passports focuses on integration with Data Repository

Service (DRS) servers, while the new Sequence Read Archive (SRA) tools offer improved flexibility and storage savings.
- Approved data access using the Data Use Ontology (DUO) involves a complex process, highlighting the importance of signing officials. There are also challenges with repackaging the Passports in Beacon, given the overly-large tokens that contain the visas.
- ELIXIR implementation utilises a three-part architecture that includes a proxy identity provider (IdP), an identity and access management system, and a Passport broker.
- The team discussed progress on developing a “Fake Passport Ecosystem” — a testing environment where data stewards can explore using Passports for data access controls.

Next steps

- Develop and circulate a survey to shape a Driver Project workshop on data access authorisation in June or July of this year.
- Progress testing scenarios of the “Fake Passport Ecosystem,” integrating various crypto techniques, visa types, and error cases.
- In conjunction with the GA4GH technical team, the group plans to move the “Fake Passport Ecosystem” to the GA4GH demo DNS namespace and AWS account, with other GA4GH demo deployed services such as htsget and DRS.

Rare Diseases Community of Interest meeting

Agenda and slides • Recording

This session featured presentations from a diverse panel of speakers, providing insights on best practices in data sharing from different perspectives within the rare disease landscape. Attendees heard data sharing use cases from the Wilhelm Foundation, the Rare Care Centre, the Undiagnosed Diseases Program Singapore, Unique, and MyGene2.

Key takeaways

- UDNI plans to host their global Undiagnosed Hackathon on 7 to 8 June, 2024, to convene clinicians, bioinformaticians, molecular biologies, and AI developers to tackle the most difficult unsolved undiagnosed diseases.
- Rare Care Centre emphasised the importance of global knowledge sharing as one of the world’s largest medical force.
- Undiagnosed Diseases Program Singapore shared the results of an analysis on clinically relevant variants from ancestrally diverse Asian genomes. The

findings helped to expand newborn screening, improve early detection of citrin deficiency, and establish an Asian-specific carrier screening panel.
- Unique emphasised that while data sharing is vital to understanding and treating rare conditions, the patient must be involved at all stages of data sharing and research
- MyGene2, a family-oriented gene discovery platform, noted that many families are willing to share their data. But there are many challenges, such as developing guidance to help families prioritise what data to share and how to share it, explaining the benefits of data sharing more clearly, and overcoming the activation energy threshold to sharing data.

Next steps

- A survey will be circulated to gain an understanding of current practices and challenges in data sharing in the rare disease domain.
    - Based on the survey, the community will form Study Groups to explore the challenges identified.
    - Outcomes of the survey will help shape a Rare Disease Community best practices toolkit, which may include:
        - direct links to GA4GH tooling and standards;
        - onboarding for newcomers;
        - potential “out-of-the-box” solutions for low income countries.

Tuesday, 23 April 2024

Unveiling GA4GH’s five-year Equity, Diversity, and Inclusion Strategic Plan

Agenda and slides • Recording

GA4GH Staff unveiled the completed GA4GH Equity, Diversity, and Inclusion (EDI) Strategic Plan. The primary objective of this plan is to establish clear goals to help GA4GH establish a more global community in a deliberate and respectful manner. The group presented specific goals from the plan, earmarked for action in 2024. Following the presentation, attendees had the opportunity to provide feedback and input during a discussion period.

Key takeaways

- The GA4GH EDI Strategic Plan is now finalised after incorporating feedback from two open for comment periods.
- The plan encompasses 21 deliverables across 3 priority areas: people, products, and practise.

- Through a prioritisation matrix exercise, GA4GH staff identified 11 deliverables as "easy wins" or projects that are already in development. During the session, attendees participated in a dot-voting exercise.
- Deep dives of implementation approaches were provided for the Best Practices and the GA4GH Diversity Survey deliverables. Regular progress updates about these and other deliverables will be communicated to the community.

Next steps

- Contributors are asked to participate in a prioritisation dot-voting exercise on Miro to help GA4GH staff focus on actioning the 11 deliverables. For questions on how to participate, please reach out to info@ga4gh.org.

Around the world in one query

Agenda and slides • Recording

Efforts within GA4GH have focused on describing genomic variations and their effects, but there is a significant gap between practical use cases and current standards in the area of variant query, especially for structural variants. To bridge that gap, attendees met to produce a non-technical roadmap for implementing variant query vocabularies.

Key takeaways

- Several needs arose, including:
    - the development of clear, conceptual definitions and standards for genomic variation queries;
    - standardisation and document normalisation strategies to ensure interoperability across resources — particularly, the normalisation of genomic data to simplify querying and ensure consistency across different platforms.

Next steps

- Develop tools and standards to manage the complexity of genomic data representation.

DRS v1.5 key features and plans for 2024

Agenda and slides • Recording

Driver Projects and other implementers reviewed the Data Repository Service (DRS) v1.5 specification to help advance the standard.

Key takeaways

- To prepare DRS v1.5, the group discussed cold storage thawing and lessons learned from dbGaP, geolocation and location constraints, reliable cloud and access method determination, DRS server metrics, and metadata linking between DRS objects and Discovery Work Stream APIs.

Next steps

- Release DRS v1.5 by GA4GH 12th Plenary in September.
- Work together on a DRS publication. If you have contributed to DRS in a tangible way, please add yourself to the authors registry.

Beacon thematic working session: cohorts

Agenda and slides • Recording

Attendees discussed requirements from Driver Projects regarding the aggregation of requests and responses within Beacon. These include returning counts by sex and age, and potential extensions to the current specification to enhance the functionality and usability of Beacon.

Key takeaways

- The meeting highlighted the integration of Beacon within various platforms, such as the International Cancer Genome Consortium (ICGC) ARGO's Ranger and McGill's Bento. These efforts aim to enhance genomic data exploration and multi-omics dataset organisation.
- The need for Beacon to better support detailed cohort descriptions, aggregated query results, and federated queries across multiple data centres emerged.
- Federated cohort building and data access, while addressing privacy concerns and standardisation of queries, were also identified as key areas of focus.

Next steps

- Enhance Beacon's cohort descriptions and aggregated queries:
    - clarify the types of cohorts (formal, Beacon-identified, or synthetic) and support them with appropriate endpoints;
    - improve Beacon's support of detailed cohort descriptions and aggregation of query results.
- Integrate Beacon with ICGC ARGO's Ranger platform and add a Beacon interface to the existing GraphQL interface.

- Enable federated search and querying across multiple databases, addressing privacy and sensitivity concerns.

Technical Assignment Subcommittee (TASC) meeting

Agenda and slides • Recording

Attendees learned about the GA4GH Technical Alignment Subcommittee (TASC), which provides community guidance for issues affecting all GA4GH standards and products, and discussed new issues.

Key takeaways

- Through an interactive whiteboarding session, attendees dove into the TASC project of moving standards into maintenance mode. The group discussed various topics, such as when and why a product should move into maintenance mode, types of maintenance, and challenges associated with reactivating contributor participation.
- To establish a uniform and consistent way of citing GA4GH specifications, TASC has settled on the solution of using digital object identifiers (DOIs). The GA4GH technical team is prioritising work to make DOI minting easier.
- Attendees discussed proposals for the namespace and resolution of GA4GH URIs issues. The proposal includes having TASC register the GA4GH namespace on identifiers.org and other identifier namespace registries, manage registration of GA4GH compact URI (CURIE) patterns used by GA4GH products, support the resolution of GA4GH CURIEs to uniform resource identifier (URI) targets, and approve the use of a persistent uniform resource locator (URL) syntax for consistency across all GA4GH schemas.

Next steps

- Develop criteria and guidance around moving GA4GH standards into maintenance mode.
- Propose an action plan for the registration and resolution of URIs for GA4GH products and data objects.

Data visiting

Agenda and slides • Recording

Attendees learned about the concept of data visiting and developed a plan of action for addressing related ethical, legal, and social implications (ELSI) for exploring the relationship between data visiting and other forms of data sharing.

Key takeaways

- Data visiting can offer numerous benefits over traditional data sharing. Despite this understanding, its use can raise significant legal and ethical concerns.
- This group will explore these challenges and develop policy tools, including a lexicon that uncovers the relationship between data visiting and other forms of data sharing.

Next steps

- Revise and circulate a data sharing lexicon within GA4GH for feedback.
- Conduct background research on data visiting.

Beacon variants

Agenda and slides • Recording

Attendees learned about example SNP/SV cases, explored extensions to the current Beacon specification to cover new use cases, and discussed typed queries (e.g. deletion, fusion, and translocation requests).

Key takeaways

- It is challenging to accurately represent complex genomic events in Beacon, and there is a need for standardised vocabulary and annotations to enhance interoperability across databases and ensure consistency in querying genomic data.
- There is a call for continued collaboration to refine standards and address challenges in genomic data sharing. The Beacon Scouts' broader role in community standard-setting was highlighted.

Next steps

- Refine the Beacon API to accommodate complex queries and extend it to support additional query features like translocations and fusions.
- Document standardised vocabulary and definitions for structural variations in genomics, with a focus on annotating variants with clinical impact and frequency information.
- Harmonise annotations across databases and explore ways to represent non-SNP variations comprehensively in data formats like VCF.

GKS: the feature is not enough

Agenda and slides • Recording

Attendees discussed a feature-centric JSON schema for sequence annotations obtained from Ensembl and RefSeq, and discussed a path to follow up on sequence-centric schema.

Key takeaways

- The team discussed the development of a feature-centric JSON schema.

Next steps

- Develop a JSON Schema for the Sequence Annotation (SA) specification, referencing existing data models for comparison.
- Define features anchored to sequence location rather than sequences themselves to ensure clarity and usability in genomic annotations.
- Clarify sequence ontology references within the schema and make a decision on the inclusion of self-defined value sets versus established ontologies.
- Delineate essential components and optional extensions in the core schema.

Wednesday, 24 April 2024

Implementation of Beacon in cancer use cases

Agenda and slides • Recording

Following up on a previous Cancer Community meeting exploring the current status of Beacon within cancer use cases, Jordi Rambla addressed community feedback and discussion points regarding changes and additions to be made to Beacon to allow for expanded utility in cancer contexts.

Key takeaways

- There is a need to expand upon the 1+MG Minimal Dataset for Cancer and other cancer data models, integrate clearer and more specific definitions of concepts and alternative solutions, and improve the accuracy of cancer data queries.
- Many different approaches were discussed during the session, including leveraging the expertise and relevant efforts across GA4GH, as well as the integration of ontologies and large language models in Beacon.

Next steps

- A follow-up community meeting will be organised with relevant Work Streams and Cancer-focused Driver Projects to address common challenges identified within Cancer Beacon queries. If there is sufficient interest, commitment, and a defined scope, a Study Group may form.

GKS: you only implement twice

Agenda and slides • Recording

Attendees learned about Genomic Knowledge Standards (GKS) product applications in production systems and explored reference implementations.

Key takeaways

- The meeting focused on ongoing implementation work within the GKS Work Stream, featuring presentations from various projects such as ClinVar, MaveDB, and BRCA Exchange.
- Efforts to standardise data formats, handle missing data within Variation Representation Specification (VRS) objects, and explore unified approaches to variant nomenclature were highlighted.
- Main challenges include ensuring interoperability, managing data dependencies, and aligning with evolving standards like the Variant Annotation (VA) specification and VRS.

Next steps

- Documentation will be developed outlining how to identify and handle missing data within VRS objects. This will help ensure consistency and clarity in handling missing data, and thereby improve interoperability and usability of VRS objects.
- The most effective approach will be determined for representing VRS alleles to accommodate diverse data structures while maintaining compatibility with consumer expectations.
- The Work Steam will discuss the creation of a unified package for variant nomenclature handling. Variant nomenclature handling will be streamlined to improve efficiency and consistency across different data sources and platforms.

Beacon resolved: a session for development

Agenda and slides • Recording

Attendees worked through the GitHub backlog, classifying issues, building consensus on OR boolean operator necessity and how to add it in a simple way, and reviewing current filter solutions.

Key takeaways

- There is an urgent need to address the challenge of harmonising responses within the Beacon network to enable aggregated information while maintaining confidentiality of response origin.
- Structuring Beacon network responses emerged as a key focus. Discussions centred on how to effectively incorporate network member information and proposals for flexible network configurations accommodating various architectures.
- The group also discussed the complexities surrounding the maintenance of an open API specification.

Next steps

- Address the challenge of current Beacon specification not supporting aggregated information from multiple Beacon responses.
- Discuss and finalise the structure of a Beacon network response, including the incorporation of network member information and a Beacon ID to identify the source.
- Evaluate the necessity of maintaining an open API in Beacon v2.0, due to the complexity of schemas involved. Consider integrating an open API as part of the protocol itself, possibly leading to the development of Beacon v3.0.

AAI & Passports “what’s next?” unconference sessions

Agenda and slides • Recording

The Data Security and Cloud Work Streams discussed various topics related to Passports, using an unconference style approach.

Key takeaways

- Considerations were made around downsizing Passport tokens and reducing permission footprints, while acknowledging the challenges related to combining multiple authorising entities and task-specific token specificity.
- While there are advantages of OID4VC Passports, as compared to GA4GH Passports, the implementation challenges of OID4VC warrant further discussion.

Next Steps:

- Continue discussions during upcoming Passports/AAI meetings through a series of presentations on these topics and use cases.

Crypt4GH, workflows, and key management

Agenda and slides • Recording

Crypt4gh is an encryption standard for securely storing data at rest. Its main advantage, compared to earlier encryption formats like OpenPGP, is offering the ability to randomly access encrypted files. One aspect left out of the specification's initial scope was key management; however, getting this right is vital to ensuring that data is kept secure.

Attendees learned about the Crypt4gh file encryption standard and discussed issues around key management, authorisation, and authentication, especially in the context of interoperating with other GA4GH standards like the Data Repository Service (DRS), htsget, and AAI/Passports. The group discussed plans to help make Crypt4gh easier to use in practice.

Key takeaways

- The team discussed specification updates, including:
  - use of additional data to enable whole-file integrity checks;
  - removal of support for multiple writers in the header;
  - support for detached headers, which reference a separate data block, for greater flexibility.

Next steps

- Integrate with other GA4GH standards.
  - Look into and assist with htsget implementations serving Crypt4gh data.
  - Work with Cloud Work Stream teams on using Crypt4gh with DRS, Task Execution Service (TES), and Workflow Execution Service (WES).
  - Provide guidance on the need for authentication, authorisation, and key management.

Federated cohort building

Agenda and slides • Recording

Attendees learned about the concept of federated cohort building, prior GA4GH work in the area, and related real-world use cases and projects. The group also discussed the need for a specification and process for federated cohort building, leveraging available GA4GH standards.

Key takeaways

- The meeting highlighted the importance of discovering relevant data before requesting access and building cohorts across multiple repositories. This

underscores the significance of enabling researchers to identify individuals with specific characteristics across various datasets for research purposes.
- Diverse use cases were presented, illustrating the challenges and objectives of harmonising data across multiple studies within large integrated discovery programs and transferring synthetic cohorts between different systems.
- There is a need to assess existing GA4GH tools and standards to address challenges related to federated data retrieval and transformation.

Next steps

- Identify specific topics of interest for deeper exploration, such as applying Beacon cohorts to asthma use cases. The team will gather requirements and use cases from the community to guide the next steps.
- Address syntax and semantics issues and agree on terminology to clarify how data is described and interpreted across different platforms and systems.
- Explore practical applications of existing tools, rather than reinvent them. This includes understanding the capabilities of different tools like Beacon and Data Connect in addressing specific data needs and leveraging them effectively to solve problems.

GKS: from the roadmap with love

Agenda and slides • Recording

Attendees recapped key points from the other GKS sessions at the Connect 2024 meeting and developed plans for the 2024-25 road map.

Key takeaways

- The group identified critical needs in the GKS Work Stream, including how to bolster the effectiveness and widespread adoption of GKS standards.
- Prioritisation to address these identified gaps are forthcoming in the 2024-2025 road map.

Next steps

- Conduct an inventory to identify and prioritise classes ready for trial use.
- Develop educational materials and support resources, particularly for smaller groups lacking engineering resources, to support implementation of GKS standards. Ideas include hosting webinars and recording introductory videos to help increase visibility and understanding of the specifications.
- Advance the maturity of downstream products by formally defining execution methods, forms, and communications, similar to processes from other standards development organisations such as HL7 and W3C.

- Prioritise the development and integration of various features and data structures related to genetic variants and their pathogenicity, as well as addressing challenges in handling copy numbers in resources like ClinVar.
- Prioritise "VRS in a Box" for further development, as it received positive feedback. Gather more input through a Google form.
- Explore the creation of a dashboard to improve visibility into the status of specifications, such as distinguishing between draft, trial use, and normative stages.

Beacon filter solutions

Agenda and slides • Recording

Attendees reviewed and resolved all issues related to filters in the Beacon Github issues backlog, produced a document proposing changes to the specification with regards to filters, and tested proposals by implementing the proposed solutions.

Key takeaways

- Filter queries in Beacon APIs need to be extended due to increased complexity and the requirements of initiatives like Matchmaker Exchange and Beacon networks.
- Finding a solution will require careful consideration, potential tooling, and further experimentation.
- Attendees discussed various solutions for extended filter logic, the intent of filter standards, and whether Beacon APIs should support logical expressions in filters.
- Comprehensive query-building tools and parsers for logical expressions are key to make data more discoverable and human readable.

Next steps

- Several approaches to filters were presented during the session. The requirements of each proposal will be reviewed and discussed in further detail at the next Beacon Scouts meeting.

DaMaSC Schema Registry

Agenda and slides • Recording

Attendees discussed the challenges and potential solutions for developing a schema registry, which would enhance consistency, clarity, and interoperability across GA4GH. Attendees heard use cases from the Cancer Research Data Commons (CRDC) and

dbGaP. There was also an exploration into how best to support consortiums in prototyping metadata for novel scientific experiments. To help inform the GA4GH Schema Registry project, the group then reviewed examples of current schema registries, such as Data+Bio, CDISC/CaDSR ISO 11179 repositories, SchemaBlocks, and Bioschemas.

Key takeaways

- The presenters emphasised the need for a schema registry. Such a registry can support interoperability, help harmonise schemas across multiple data commons, and recommend ways to systematically describe data pipeline inputs.
- A central repository could help identify crucial metadata elements and ensure consistency across different labs and experiments, particularly for fields such as neuroscience.
- Attendees agreed on the importance of considering a federated model for schema registries and interoperability standards to accommodate diverse needs and structures across different studies and experiments.
- The group reviewed existing resources, such as Schema Blocks, BioSchemas.org, Metadata Repositories (MDRs), and fairsharing.org.

Next steps

- Review and collect feedback on the requirements needed for the GA4GH Schema Registry.
- Develop and implement a proof-of-concept for the GA4GH Schema Registry — including a submission template and a registry landing page — for approval by the GA4GH Steering Committee.

Driver Project workshop: multi-site / geography collaboration and governance

Agenda and slides • Recording

Attendees learned about Driver Project (DP) needs, as they relate to the ethical, legal, regulatory, and technical challenges to enabling multi-site collaborations. Attendees also heard presentations from specific GA4GH teams involved in this topic, including the GDPR Forum, Ethical Provenance, Cloud standards, and Passports.

Key takeaways

- The International Precision Child Health Partnership (IPCHiP), Human Heredity and Health in Africa (H3Africa), the National Institutes of Health Cloud Platforms Interoperability (NCPI) effort, and the European Genomic Data Infrastructure

(GDI) project have unique organisational requirements within the area of multi-site collaboration as it relates to data access, governance, management, and infrastructure needs for data sharing agreements across institutions and jurisdictions.

- Collaborations and regular meetings from Work Stream leads with DPs can help to better address these needs through the development of GA4GH products.

Next steps

- Map out suggestions from the Work Streams that are most relevant for tackling the challenges outlined by the DPs.