GA4GH refget API

Minutes and Actions 2023,2024,2025,2026

These are the minutes for the refget API meetings, a subgroup of the GA4GH Large Scale Genomics Work Stream. For further information, please visit the GA4GH Work Stream page.

Let's make the world slightly more interoperable and end the chr1 vs 1 debacle of bioinformatics!

Roadmap - 2024

1) Submitting sequence collections to PRC and steering committee approval Q2 or Q3 delivery
2) Increasing adoption of sequence collections & refget including integration into 3rd party toolkits e.g. BED, CRAM
3) Pangenomes and linking up with HPP for DP support to do this
4) Alignment with VRS 2.0 - Done

Roadmap - 2025

1) Submitting sequence collections to PRC and steering committee approval Q2 or Q3 delivery - **Done**
2) Increasing adoption of sequence collections & refget including integration into 3rd party toolkits
    - BED, CRAM
    - Refgenie
    - OmniPy (Sveinung)
    - RDA Genomic Annotations
    - Ensembl and ENA services
3) Pangenomes and linking up with HPP for DP support to do this.Pangenome User stories
4) Alignment with VRS 2.0 **Done**

Priorities for 2026

- 1. pangenomes (spec development work)
- 2. practical applications
    - store/distribution (solve challenge with distribution)
    - EVA validator? (improve current tool that will validate whether variant is using a reference genome we recognize)
- 3. integrations into other standards (file formats)
- 4. deployment of services (ensembl, EVA, refgenie)

Product maintainers. Responsibilities:

- Maintains LSG product(s)
- Implements and documents approved updates to products

- May maintain inward- and outward-facing web sites, repositories, etc
- Manages products through the post-development release process
- Person could be from a central GA4GH resource medium/long term
- Short-term person would have to come from our local workstream resource

Meeting Protocols

Agendas, Minutes and Action

2023-09-12: Reverse Lookup discussion - CANCELLED
2023-09-06: Refget discussion
2023-08-23: Refget discussion
2023-07-26: Refget discussion
2023-07-18: Reverse Lookup discussion
2023-07-12: Refget discussion
2023-06-28: Refget discussion
2023-06-20: Reverse Lookup discussion
2023-06-14: Refget discussion
2023-05-31: Refget discussion
2023-05-23: Reverse Lookup discussion
2023-05-17: Refget discussion
2023-05-03: Refget discussion
2023-04-25: Reverse Lookup discussion
2023-04-19: Cancelled (due to Connect meeting)
2023-04-05: Connect prep and ADR
2023-03-28: Reverse Lookup discussion – CANCELLED
2023-03-22: Refget discussion
2023-03-08: Refget discussion
2023-02-28: Reverse Lookup discussion
2023-02-22:Refget discussion
2023-02-08: Refget discussion
2023-01-31: Reverse Lookup discussion
2023-01-25: Refget discussion
2023-01-11: Refget discussion
2023-01-03: Reverse Lookup discussion - CANCELLED
2023-MM-DD: Template

Meeting Protocols

- Please note that by participating in meetings, attendees agree to adhere to the GA4GH Standards of Professional Conduct.
- Meetings may be recorded for note-taking purposes. Recordings will be deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd
- 2022 Previous minutes available here
- 2021 Previous minutes available here
- 2020 Previous minutes available here
- 2019 Previous minutes available here
- 2018 Previous minutes available here
- https://w3id.org/ga4gh/minutes/refget should point to latest minutes

Agendas, Minutes and Action

YYYY-MM-DD: Template

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

2026-03-04:

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time

1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Lack of participation in meetings: should we move to monthly? Should we take a hiatus and come back in the fall? move to ad hoc meetings? Keep the schedule but start cancelling meetings without agendas? Set a long-term agenda and broadcast better?	
4.		
	A.O.B.	

Meeting recording:

Meeting minutes:

2026-02-18:

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

Meeting recording:

Meeting minutes:

2026-02-04:

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Raise PR for creating a prefix for sequence collections following the TASC readme	Tim	
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	RefgetStore updates, and: https://refget.databio.org/fasta	
4.		
	A.O.B.	

Meeting recording:

Meeting minutes:

2026-01-21:

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,
Apologies : Sveinung Gundersen

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Nathan will prepare demos of the Refget Store and deploy it on a demo instance to show how it works and analyze disk usage.	Nathan	
2	eva-seqcol: to be put in production	Tim	
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B. Reminders:Connect session proposals Connect proposal - Submit through Nolt - https://ga4gh-connect-session-proposals.nolt.io/ ● [Register for April Connect 2026]() ● [Learn more and submit proposals - deadline: January 30]() 📅 Proposal Timeline Opens: 1 December **Closes: 30 January** Secretariat review: 2 to 13 February Sessions confirmed: 20 February	

Meeting recording:
Meeting minutes:

2026-01-07:

Chair: ~~Andy Yates~~ / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Reggan Thomas , , , , , ,
Apologies : Sveinung Gundersen

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

1	Nathan will prepare demos of the Refget Store and deploy it on a demo instance to show how it works and analyze disk usage.	Nathan	
2	eva-seqcol: to be put in production	Tim	
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Priorities for 2026? ● 1. pangenomes (spec development work) ● 2. practical applications ○ store/distribution (solve challenge with distribution) ○ EVA validator? (improve current tool that will validate whether variant is using a reference genome we recognize) ● 3. integrations into other standards (file formats, VRS?) ● 4. deployment of services (ensembl, EVA, refgenie)	
4.		
	A.O.B. Reminders:Connect session proposals Connect proposal - Submit through Nolt - https://ga4gh-connect-session-proposals.nolt.io/ ● Register for April Connect 2026 ● Learn more and submit proposals - deadline: January 30 📅 Proposal Timeline Opens: 1 December **Closes: 30 January** Secretariat review: 2 to 13 February Sessions confirmed: 20 February	

Meeting recording:
Meeting minutes:

2025-12-10:

Chair:~~ Andy Yates ~~/ Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Sveinung Gundersen, Jinny Park

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	eva-seqcol: to be put in production	Tim	
2	Nathan will prepare demos of the Refget Store and deploy it on a demo instance to show how it works and analyze disk usage.	Nathan	
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Update on refget fasta and RefgetStore ideas	Nathan
4.		
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/RGNWM0RDQBYqNaQWDgN11oJhwJT6SZCBxpSr-zKH5xNBnxyto3C8rg6kGdN2--b6.r3u7JMmaGnj_DikR?startTime=1765378788000

Chat messages:

Nathan
https://docs.bedbase.org/gtars/refget-store-format/
https://docs.bedbase.org/gtars/python/refgetstore/

Sveinung
https://genome.crg.es/annotrieve/

Meeting minutes:

Review of Previous Actions & Context

- WGS QC Standard Inquiry:

  - Justin previously inquired about specifying how the reference genome is used in QC workflows. They were looking for a way to retrieve GRCh38 (base only) directly.
  - Timothee clarified that Refget provides definitions but not necessarily a direct download of a FASTA file in the way Justin envisioned, though Refgenie is closer to that need.
- Ensembl Sequence Collections:
  - Timothee reported no firm plans at Ensembl to implement sequence collections in the coming year.
  - While the code seems done, it has not been a high enough priority to move into production.

FASTA DRS Endpoint

- Nathan reported on his work regarding serving sequences from a "collections" perspective rather than a "sequence" perspective to address server load issues.
- Nathan finished implementing a FASTA endpoint that mounts a DRS API compatible interface.This provides access URLs to get a FASTA file representing a sequence collection.
- Serves as a mechanism for getting sequence data in the context of a collection, distinct from the Refget sequences service which handles individual sequences.
- Nathan plans to switch Refgenie to use this as its source of FASTA. He also implemented FAI tracking in the metadata.

Refget Store

- Discussion moved to a more advanced data structure Nathan developed called the "Refget Store."
- Structure: A file-based structure containing a "collection-less sequence store" (keeping all sequences) and a "collection store" (mapping names to sequences).
  - Allows retrieving subsequences from the context of a collection without duplicating sequences.
  - Encodes sequences using alphabet-specific bit-packing for compression.
  - Can be hosted on an HTTP service (e.g., S3); a local client can dynamically cache only the necessary sequences.
- Nathan views this as how the sequences standard "should have worked," utilizing whole sequence file pulls and local substringing rather than remote substringing.

Implementation & Use Cases Discussion

- The group debated the merits of the simple FASTA endpoint versus the more complex Refget Store.
- Data Duplication vs. Dynamic Construction:
  - Timothee raised the issue of multiple naming conventions requiring duplicated storage if static FASTA files are used.

  - The Refget Store solves this by separating sequence data from naming conventions, allowing dynamic FASTA construction without data duplication.
- Sveinung noted this would be excellent for Galaxy, allowing them to distribute the structure and use a small tool to extract exact files needed for jobs.
- Timothee noted that large organizations have inertia and existing infrastructure that makes adopting new storage methods difficult, though he sees the value in the efficiency.
- The FASTA DRS endpoint is useful for simple users.
- The Refget Store is a powerful tool for power users and enables the creation of FASTA files on the fly.
- Nathan suggests the Refget Store might supersede the need for a standalone FASTA endpoint because a client can easily generate the FASTA from the store.

Other Business

- Sveinung introduced Emilio Righi, who developed a tool for downloading GFF files.The tool calculates statistics and compares species, which aligns well with sequence collections.
- Jinny will leave a note for Reggan to add Emilio to these meetings
- Next meeting will be in January (Dec 24 meeting cancelled)

Actions

- Nathan will prepare demos of the Refget Store and deploy it on a demo instance to show how it works and analyze disk usage.

2025-11-26: Integrate Refget

Chair: ~~Andy Yates / Nathan Sheffield /~~ Tim Cezard
Attendees “Name (Affiliation)”: , Justin Jeyakani (PRECISE Singapore), Reggan Thomas (GA4GH), , , , ,
Excused: Sveinung

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	

3.	Support integrate Refget seq into WGS QC	Justin / 10 mins
4.		
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/xsorLnaWLiSoQtShFq5c-FjwlR-PidwJY9OZZxdbf4a0dUdf3xM0U1rcDcbjMAB-.kp12HbGxPxNwlKrc

Chat messages:
TC - https://ui.refgenie.org/genomes
TC - https://refgenie.org/refget/notebooks/digests/

Meeting minutes:

Justin from Singapore, part of the WGS QC standard development team, discussed the integration of ref-seq API for querying genome builds instead of linking directly to specific repositories like the 1000 Genome Project. Tim Cezard explained the ref-seq API, which includes ref-seq sequences and ref-seq sequence collections, and how it can uniquely identify sequences using MD5 or GH4 digest algorithms. Justin's team currently validates MD5 sums but is considering using ref-seq collections for better standardization. They also discussed the potential use of ref-seq collections in pangenome identifiers and the need for more community engagement and user stories to refine the standards.

Discussion on Refget API Integration

•Justin inquires about using the Refget API to query genome builds instead of linking directly to specific genomic repositories.
•Tim Cezard explains the two parts of Refget: Refget sequences and Refget sequence collections.
•Refget sequences identify single sequences using MD5 sums, while Refget sequence collections identify groups of sequences using a two-step construction of final identifiers.
•Tim Cezard provides details on the different digest algorithms used and the features of Refget sequence collections.

Implementation and Use Cases
•Justin asks about the practical implementation of using Refget sequence collections in their workflow.
•Tim Cezard explains that different tools will have different implementations and that Refget sequence collections can be used to define specific genomes.

•Justin mentions the need for specific naming conventions for targeted regions and the compatibility with Refget sequence collections.
•Tim Cezard discusses the use of Refget sequence collections in tools like Ref Genie for managing genome assets and indexes.

Technical Support and Next Steps
•Justin seeks technical support for integrating Refget sequence collections into their workflow.
•Tim Cezard suggests checking with Nathan for downloading whole genome builds and mentions the use of Ref Genie for managing genome assets.
•Justin plans to explore the feasibility of using Refget sequence collections and validate the MD5 sums.
•Tim Cezard and Justin discuss the potential benefits of using Refgett sequence collections and the need for further exploration and validation.

Meeting Conclusion and Future Plans
•Reggan mentions the next meeting scheduled for December 10 and the possibility of canceling the December 24 meeting.
•Justin plans to join the next meeting and possibly communicate with Nathan for more detailed information.
•Tim Cezard and Reggan discuss the focus on pangenome identifiers for 2026 and the need for more community involvement and user stories.
•The meeting concludes with plans to send an email about the upcoming discussion topics and to wait for further input from Nathan.

Action Items
- Follow up with Nathan to understand how to retrieve the GRCh38 FASTA file using the ref-Genie tool.
- Explore the feasibility of using the refget sequence collection standard in the participant's workflows and pipelines.
- Provide an update on the discussion and any progress in the next meeting scheduled for December 10th.

2025-11-11: eva-seqcol & Github review

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Nathan, , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	eva-seqcol: to be put in production check Ensembl plans	Tim	
2			

3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda - Report from GA4GH Plenary Uppsala	
2.	Previous Actions Review	
3.	Check in on implementation in progress - eva-seqcol: to be put in production - check Ensembl plans	
4.	GitHub issue review: Spend 15 minutes reviewing and (hopefully) closing old issues to clean up the issue tracker. Also make an initial response to new issues and PRs.	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/N540Eect4RR-veuKM_Wt0YHy8HesencO2kIgDZjs5TDZy1L9d6ws3f3lqGz7wV77.2d6zmEPuPjfV41D7
Meeting minutes:

2025-10-15: Github review

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Nathan, Tim (briefly), , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time

1.	Welcome, Introductions & Confirm Agenda - Report from GA4GH Plenary Uppsala	
2.	Previous Actions Review - [ ] Get final approval on user stories document	
3.	GitHub issue review: Spend 15 minutes reviewing and (hopefully) closing old issues to clean up the issue tracker. Also make an initial response to new issues and PRs. - https://github.com/ga4gh/refget/pull/108 - https://github.com/ga4gh/refget/pull/109 - https://github.com/ga4gh/refget/pull/111	
4.		
	A.O.B.	

Meeting recording:
Meeting minutes:
- Meeting ended after brief discussion due to lack of attendees.

2025-10-01: Refget Pangenomes

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Read through and approve new User Stores document for posting: https://docs.google.com/document/d/1dnW9wXGcWDzoyLXX8cWfsL1xTMJ8w6DX_qdSoV51eBg/edit?tab=t.0#heading=h.f6fcw4vffacl	All	2 weeks
2	Close https://github.com/ga4gh/refget/issues/101	Nathan	PR Opened.
3	Follow up with TASC issue #59 about the use of prefix in sequence collection level1	Tim	

	Agenda Item	Person/Time

1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	GitHub issue review: Spend 15 minutes reviewing and (hopefully) closing old issues to clean up the issue tracker. Also make an initial response to new issues and PRs. - https://github.com/ga4gh/refget/pull/105 - User stories: https://docs.google.com/document/d/1dnW9wXGcWDzoyLXX8cWfsL1xTMJ8w6DX_qdSoV51eBg/edit?tab=t.0#heading=h.f6fcw4vffacl	15
4.	Graph representations and refget pangenomes https://github.com/ga4gh/refget/issues/103	
5.	Dealing with pangenome multi-level names https://github.com/ga4gh/refget/issues/104	
	A.O.B. Next meeting - 15 October at 3 pm BST/10 am ET /7 am PT	

Meeting recording:
Meeting minutes:
https://us02web.zoom.us/rec/share/NalagK2cJW0R2PW64gMCJ29C7yTRc2vSet9L_fn1_xih4duR4dFZDeSyvyyOmriA.Nfjg5wX8CLu76_Le

2025-09-16: Pan-genome identifiers

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Reach out to TASC re: prefixes	Tim/Andy	
2	Review the poster - Refget section - ***Building the Future of Reference data – Sequence collections***	Nathan/Tim/Sveinung	25th Sep
3			

	Agenda Item	Person/Time

1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Github issues review	
4.	Discuss the representation of pan-genome identifiers, including whether the graph structure should be inherent or represented as a separate level one attribute	
	A.O.B. LSG poster - [Refget section](#) - review Deadline - Sep 25, 2025	Reggan

Meeting recording:
Meeting minutes:
TC - https://github.com/samtools/hts-specs/issues/824

2025-09-03: Prefix for Sequence Collection

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Update on paper status	Nathan

4.	Opening Session at GA4GH Connect (**9:00 AM CEST, 8 Oct**). Would you like to feature anything—like the Sequence Collection manuscript, Pangenome Identifier use cases, or other updates? If so, please share by **ASAP**, including: A brief description of the update/product Presenter name Please add it here - If andy or sveinung are available, they could share: 1. the paper; 2. the SCOM: https://refget.databio.org/scom as an implementation.	Reggan
5.	Review of Prefix for Sequence Collection: #53 Proposal by TASC for defining prefixes: ● Prefixes SHOULD be short, ideally 2-4 characters. ● Characters MUST be uppercase only. ● Prefixes SHOULD be for concrete types, not polymorphic parent classes. ● A prefix MUST map 1:1 with a schema type. Should we have 1 prefix for sequence collection? eg SC	Tim
6	Github issues review	
7	Discuss the representation of pan-genome identifiers, including whether the graph structure should be inherent or represented as a separate level one attribute	
	A.O.B.	

Meeting recording:
Meeting minutes;

2025-08-20: Meeting Free Week

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

Meeting recording:
Meeting minutes

2025-07-09: Pangenome user stories

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Scheduling meetings: ● July 23rd - Cancelled ● August 6th - Cancelled ● August 20th - Meeting free week but can still catch up if necessary	Tim
4.	Pangenome User stories	

	A.O.B.	

Meeting recording:
Meeting minutes:

2025-06-25: Sequence collection integration in VCF

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Consider submitting a proposal for a presentation at the upcoming Connect session.	all	Jul 3, 2025
2	Determine the appropriate 2-letter prefix for sequence collection identifiers (e.g., SC) and confirm with the tasc group		
3	Discuss the representation of pan-genome identifiers, including whether the graph structure should be inherent or represented as a separate level one attribute, with Nathan.	Nathan	Next meeting

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Sequence collection integration in VCF: https://github.com/samtools/hts-specs/pull/832 ● We will need guidance from TASC on what we should use as an entity type name (SC ?) see https://github.com/ga4gh/TASC/issues/16 ● Should we allow/encourage the use of level1 identifiers? If yes they should have different types. ● ⇒ For level1 we could keep type=SC and add the property name before the identifier: ga4gh:SC.names.gklepkjfnlrghjcbhjslv	Tim
4.	Discussion topic: is the graph an inherent part of the pangenome? or is it just the sequences?	
	A.O.B. Connect session proposals - July 3rd. Submit your proposal here	

	GA4GH 13th Plenary in Uppsala, Sweden (October 6 - 10) Registration is open. Detail can be found here - https://broadinstitute.swoogo.com/ga4gh13plenary/7101555 Active contributors receive a discounted rate by taking this survey - https://ga4gh.fillout.com/t/4cC2PkNva4us	

Meeting recording:
https://us02web.zoom.us/rec/share/pH_AmG1PaoGrn4aABclEEFEvaw1xgp0tN1SswKrtgZSjsA6ZBof-Ydc_FfaKOROo.NQmTbynXWTwc858H

Meeting minutes:

Action items from the file formats meeting attended by Refget team:

SL No	Actions Arising	Assigned To
1.	Raise an issue or PR on the HTS specs for incorporating sequence collection identifiers in the SAM header.	?
2.	Make additional notes on the questions and discussion points from the meeting for the SAM header changes.	Everyone
3.	Provide a list of pros and cons for the sequence collection changes, and link to it from the HTS specs issues	Nathan
4.	Determine if sequence collection identifiers need to be namespaced, and if so, decide on the namespace	Discuss with TASC
5.	Raise an issue or PR on the VCF specs for adding a sequence collection header tag.	Done

TC : https://github.com/samtools/hts-specs/issues/824

Meeting summary:
The meeting discussed the introduction of sequence collection in VCF headers, proposing a structured identifier format. Andy highlighted the need to resolve sequence collection identifiers

using the GA4GH namespace. The team debated the inclusion of graph identifiers in Pan genome representations, with differing opinions on whether to include them at level one or level zero. They also reviewed the upcoming Connect session, noting the deadline for proposal submissions on July 3rd and the importance of cross-stream collaboration with GA4GH, and Elixir. The meeting concluded with a reminder of the next few meetings and the need for further discussions on identifiers and namespaces.

Discussion on PR for Sequence Collection in VCF Header

•The main topic is a PR on SG spec to introduce sequence collection in VCF header.

• Andy is asked about the type of task for the sequence production entity and the ongoing discussion on task.

•The need to add a summary on the issue and the developing dependency in sequence collections is discussed.

In-Person Test Meeting and TASC Management

•An in-person test meeting is scheduled for July 16th and 17th, with the main working days being the 16th and 17th.

•The discussion includes the need to figure out the two letters for the tasc and the importance of attending the meeting.

•The conversation shifts to the level one identifiers and the potential addition of structured identifiers in the PR.

•The need for a different type to distinguish level one from level zero is discussed.

Resolution and Namespace for Sequence Collections

•The conversation delves into the resolution of level one identifiers and the potential use of the C prefix.

•The discussion includes the importance of having a resolver for the GA4GH namespace and the utility of resolvable identifiers.

•The need for a structured format in VCF files is highlighted, with options for unstructured and structured formats.

•The importance of having a resolver for sequence collections and the potential issues with unique identifiers is discussed.

Integration of Sequence Collections in VCF and Other Formats

•The discussion includes the integration of sequence collections in VCF and the challenges with other formats like SAM, BAM, and CRAM.

- The need to add sequence collection identifiers to VCF first due to its flexibility is emphasized.
- The pros and cons of sequence collection changes are briefly mentioned, with a focus on pushing for VCF integration.
- The importance of having a resolver for sequence collections and the potential issues with unique identifiers is reiterated.

Pan Genome Identifiers and Graph Representation

- The discussion shifts to the identifying elements of Pangenome and the importance of the source of data.
- The potential inclusion of graph representation as an identifying element is debated.
- The need for a standardized way to include graph identifiers in Pan genome representations is discussed.
- The conversation includes the importance of having identifiers for the source and the graph to ensure compatibility and differentiation.

Connect Session and Proposal Submission

- The Connect session is discussed, with a deadline for submission on July 3rd.
- The platform for collecting proposals is introduced, and the advantages of seeing other work streams' proposals are highlighted.
- The discussion includes the potential for proposing something on Pan genome and the importance of collaboration with other work streams.
- The need for a meeting to find synergies between different work streams and the importance of active participation in the Connect session is emphasized.

Final Remarks and Next Steps

- The meeting concludes with a reminder of the next couple of meetings due to summer holidays.
- The importance of continuing the discussion on identifiers and namespaces in the next meeting is mentioned.
- The meeting ends with a note on the importance of finding synergies and collaborating with other work streams.

2025-06-11: Seqcol identifiers

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Updates on reference genome comparison analysis	Nathan
4.		
	A.O.B. There are still several problems with links on the ga4gh product page: https://www.ga4gh.org/product/refget/ - Tools and Platforms link says GitHub but it actually goes to the spec. This should go to the github repo. - Latest Version link should go to the spec, but instead it goes to the old docs - "Use this product" link also still goes to the old docs.	

Meeting recording:
Meeting minutes:

Enumerated list of why it makes sense to include seqcol identifiers in SAM/BAM/CRAM/VCF headers:

- **Provenance** benefits of using collections instead of sequences:
    - Sequences may not represent the entire genome, but only the ones relevant to the content of the file; this means the sequences alone couldn't be used to reproduce the file, but the original genome could.
    - Names may be excluded from the sequences, but will be built-in for collections
- **File size** benefits:
    - It's more compact than including sequence headers. This can be substantial for transcriptomes, not-fully-assembled genomes, or any collection with thousands of sequences.
    - It reduces duplication across files. Generally, in a set of files, each will have the same sequence set. If using sequences, this information is duplicated across all files. With collections, you eliminate that duplication.
- **Compatibility**: It makes it easier to check if two bam files are compatible. It enables compatibility check across file types, through the refget seqcol comparison function. For example between bam and cram or bam and vcf.

- **Annotation offloading**: It makes annotating the sequence collection independent from the files; annotations happen at a federated provider level. (for example, seqnames, or topologies, or whatever, are no longer duplicated in every bam file)
- **Conceptual alignment.** It aligns better with the way we actually conceptualize data analysis; a bam file really is a downstream asset of some analysis referencing a genome -- not an independent set of sequences that can change with each file. these genomes are re-used frequently as such, not as independent collections of sequences.
- **Fetch efficiency**: Can be helpful for CRAM by reducing the fetch burden; fetch once for an entire collection instead of once per sequence.
- **Promotes use of digested reference genome identifiers**. If in use in BAM, will increase uptake in other areas, which will improve reproducibility.
    - Also increases interpretability of the bam file; That thing I have in my bam header, I can paste it into the seqcol server and learn a lot about it.

2025-05-28: Update on refget implementation

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	See if we can hijack the June 17th File Formats meeting	Reggan	Done.
2	Refget contributors read/comment on this issue: https://github.com/samtools/hts-specs/issues/824	Andy/Sveinung/Tim/Nathan	Before June 17th meeting
3.	Discussion topic for next time: is the graph an inherent part of the pangenome? or is it just the sequences?		Added to agenda.

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Small update on local refget implementation	
4.	Pangenome User stories	
	A.O.B. ● Formatting, spelling, and minor wording improvements to the Refget Sequences spec (PR #102)	

Meeting recording:
https://us02web.zoom.us/rec/share/wdYfGbv4Ilbtr8q93ajbTJKdv29ikQNljcUf7UBnE1A5kUi6Xkz1G4o2VWQldVau.nOfK-p7glWP4H2TR

Meeting minutes:

The meeting discussed John Marshall's minor improvements to the sequences specification, which were deemed minor and did not require a version bump. Nathan Sheffield proposed merging the changes. The agenda included reviewing previous actions, pan genome user stories, and a local ref get implementation update. Nathan presented a Rust-based local sequence cache system optimized for memory and disk usage. The team debated the inclusion of sequence collection identifiers and names in pan genome identifiers, emphasizing the need for reproducibility and downstream compatibility. They also discussed scheduling a meeting with Rob to discuss file formats and the potential use of a DRS endpoint for sequence collections

Review of Minor Improvements to Specification

- Nathan Sheffield mentions Andy wrote the spec and should review the changes.
- Tim Cezard recalls Andy was busy and might need more time to review.
- Nathan Sheffield asks Andy to confirm if the changes are fine for merging.
- Andy confirms he will check the changes and decide on merging.

Meeting Agenda and Action Items

- Nathan Sheffield outlines the meeting agenda: previous actions, pan genome, user stories, and a local ref get implementation update.
- Nathan Sheffield reviews action items from the last meeting, including Reagan reaching out to Rob.
- The group discusses scheduling a meeting with Rob and the file formats meeting on June 17th.

Discussion on File Formats Meeting and Action Items

- Nathan Sheffield suggests hijacking the file formats meeting to discuss ref get file formats.
- Reggan agrees to check with James Bondfield about hijacking the meeting.
- Nathan Sheffield lists other action items, including opening an issue on the ref get spec and discussing a Durst-based system for accessing reference sequences with Rob.
- Tim Cezard and Nathan Sheffield discuss the separation of sequence collection inclusion in file formats and the CRAM registry issue.

Local Ref Get Implementation Update

- Nathan Sheffield provides an update on a local ref get implementation in Rust.
- The implementation allows ingesting a FASTA file, loading sequences into memory, and writing to disk.
- Nathan Sheffield explains the benefits of the implementation, including fast local lookups and reduced memory and disk usage.
- The group discusses the potential use of the implementation for CRAM and other sequence lookups.

Pan Genome User Stories and Discussion

- Tim Cezard presents user stories for resolving a pan genome and calculating a pan genome identifier.
- Nathan Sheffield explains his implementation of pan genomes in the ref get package.
- The group discusses the inclusion of sequence collection identifiers and sequence identifiers in pan genome calculations.
- Nathan Sheffield emphasizes the importance of including names in the identifiers for reproducibility and downstream compatibility.

Graph Construction and Identifier Discussion

- The group discusses the construction of graphs for pan genomes and the impact of different naming conventions.
- Nathan Sheffield argues for keeping the pan genome and graph as separate entities.
- Tim Cezard suggests considering the graph as part of the pan genome identifier.
- The group agrees to continue the discussion in the next meeting, focusing on the relationship between pan genomes and graphs.

2025-04-30: CRAM / refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,James Bonfield ,Sveinung Gundersen ,Reggan Thomas , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Reach out to Rob to discuss adding support for sequence collections in HTS-lib.	Reggan	
2	Open an issue on the hts spec to discuss adding support for sequence collections in the headers	James	Done
3	Discuss the idea of a DRS-based system for accessing reference sequences with Rob.		Working on setting up this meeting.

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	CRAM / refget / file formats discussion (?)	
4.	Pangenome User stories	

	A.O.B. Meeting on May 14, 2025 has been cancelled Next meeting is on May 28, 2025 Link to the google folder - Refget API	

Meeting recording:
https://us02web.zoom.us/rec/share/LeOYkMxVUWkA13fuhQ3HQLpbkXbF5aYqL0UlPRZiROUmbD6sbtQ0ZV1vpAsGO5XT.I9b8AXMhQiWsgQr5

Meeting minutes:
The meeting discussed the challenges and potential solutions for managing sequence collections and file formats, particularly CRAM. James highlighted the need for local caching to alleviate the load on EBI's reference registry. Nathan Sheffield proposed adding a DRS endpoint to sequence collections to provide gzipped FASTA files, reducing API requests. Tim Cezard emphasized the importance of sequence collection headers for data provenance. The team agreed to discuss these proposals with Rob and to update the CRAM spec to support sequence collections. They also planned to address pan-genome discussions in future meetings.

Action items:

- Reach out to Rob to discuss adding support for sequence collections in HTS-lib.
- Open an issue on the ref-get spec to discuss adding support for sequence collections in the headers
- Discuss the idea of a DRS-based system for accessing reference sequences with Rob.

Caching and Sequence Collections
James explains the caching layer at Sanger, which acts as a permanent squid proxy.
Nathan Sheffield and James discuss the efficiency of caching and the potential benefits of sequence collections.
James suggests that users should have local caches to solve the issue of the EBI reference registry.
Nathan Sheffield proposes adding a drs endpoint to sequence collections to provide gzipped FASTA files, reducing the load on the server.

Discussion on File Compression and Sequence Downloads
Tim Cezard questions how sequences are compressed in practice.
James explains that sequences are not always downloaded compressed and that local caches are common.
Nathan Sheffield suggests that the EBI ref get sequences are never served compressed, which is part of the problem.

Challenges with EBI Refget Server

Tim Cezard and James discuss the limitations of the EBI refget server and the need for a copy in SV.

James mentions that most of the ref get bandwidth is from one American University, indicating a management problem.

Nathan Sheffield and James discuss the need for user education and proper configuration to avoid denial of service at EBI.

James suggests removing the fallback to EBI to encourage users to specify their reference paths.

Proposed Solutions and Next Steps

Nathan Sheffield presents an alternative perspective, suggesting a CDN-first approach for sequence data.

James agrees that caching is the solution but notes that users are often lazy in implementing it.

Nathan Sheffield proposes adding sequence collection information to CRAM headers to prioritize caching.

Tim Cezard and James discuss the potential benefits and challenges of implementing sequence collections in CRAM headers.

Sequence Collection and Data Provenance

Tim Cezard suggests adding sequence collection information to CRAM headers for data provenance.

James expresses support for data provenance but notes the challenges of implementing it.

Nathan Sheffield proposes embedding sequence collection information in BAM files to avoid reliance on external services.

James and Nathan Sheffield discuss the potential benefits and drawbacks of different approaches to sequence collection integration.

Digital Object Reference System (DOR) and File Access

Nathan Sheffield proposes using DOR to provide file-based access to sequences, enabling a CDN-like approach.

Tim Cezard and James discuss the need to define this approach in the spec and get implementation support.

Nathan Sheffield suggests discussing this with Rob to get his input on the feasibility and benefits of the proposed approach.

Tim Cezard and James agree on the importance of defining the approach and getting community support.

Next Steps and Action Items:

Nathan Sheffield summarizes the next steps, including discussing with Rob and defining the sequence collection spec.

Tim Cezard suggests opening an issue on the HDS spec to track the discussion and ensure it doesn't get forgotten.

James agrees to report back to Rob and discuss the proposed approach in the next release of HDS lib.

Nathan Sheffield and Tim Cezard discuss the importance of addressing the issues with EBI ref get server and improving data provenance in CRAM files.

2025-04-16: Connect key aways and Issue 88

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Sveinung Gundersen ,Reggan Thomas , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Start a pangenome use cases document?		
2	Continue working on paper	All	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	5 mins
2.	Previous Actions Review	5 mins
3.	Connect review. - Approve Key Takeaways and Next Actions (below) - Should we talk about the sequences bottlenecks?	30 mins
4.	Adoption: Go over status of current services, what's needed to drive implementation	10 mins
5.	https://github.com/ga4gh/refget/issues/88	10 mins
	A.O.B. Next meeting during **GA4GH Meeting Free Week**. Cancel / hold / reschedule ? Key takeaways and action items for October Connect 2025: W GA4GH Connect April 2025 - Refget Sequence Collections Standards …. ☐ Key Takeaways ☐ A sequence selection standard has been introduced to uniquely identify and compare collections of sequences. ☐ The standard includes a schema, digest algorithm, and API to support validation and comparison. ☐ Implementations are already in use (e.g., refget, EVA) for tasks like validating VCFs and harmonizing metadata.	

	☐ The approach may extend to pangenomes, offering recursive identifiers for complex genome representations. ☐ Broader infrastructure support (e.g., public servers, CDNs) and community adoption are essential for scale and sustainability ☐ Next Actions ☐ Encourage more institutions to host refget-compatible servers to improve reliability and access. ☐ Explore CDN-based distribution of sequence collections for scalability. ☐ Continue refining pangenome identifier strategies. ☐ Engage with large-scale genome providers (e.g., Ensembl, NCBI) for implementation support.	

Meeting recording:
https://us02web.zoom.us/rec/play/KL71joOs63TKsJ99ZBvJWTwQnSVPndpWkUqJHYzzBHt3udJUaN1Rav5944PWP-gIQpCCI3C5iRxWrNyI.yVFV-ZVoIphTtIec?accessLevel=meeting&canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2Fva-plyIvsQz8iKJBZ9S3TPS3M5E401tprRjKEb1uITUiT7LiAiiaNnV5QgVwD4Fc.nSzrPCuir50o8izP

Meeting minutes

Review of Previous Actions and Connect Takeaways

Nathan Sheffield reviews previous actions, including slides, paper review, and digest computations.

Tim Cezard mentions an issue with name length pairs resolved by Tim and Sven.

Nathan Sheffield and Tim Cezard discuss the paper, with Tim planning to write an implementation description and use case.

Nathan Sheffield emphasizes the importance of comparing discoverability and compatibility in the paper.

Key takeaways from Connect include the introduction of the sequence collection standard and the need for broader infrastructure support.

Discussion on Sequence Collection and Pan Genome Integration

Tim Cezard plans to ingest large numbers of genomes and push the resulting Docker image to Docker Hub.

Nathan Sheffield suggests using Docker Compose for multiple services, but Tim prefers loading data outside the Docker file.

Nathan Sheffield and Tim Cezard discuss the efficiency of using compressed data for sequence collections.

Nathan Sheffield proposes a Docker endpoint for downloading FASTA files, which Sveinung Gundersen supports as a solution for Galaxy infrastructure.

Andy Yates explains the challenges with ENA's current infrastructure and the need to move away from Oracle.

Ref Get Sequences and CRAM Format Challenges

Nathan Sheffield proposes using sequence collections for CRAM files to improve efficiency.

Andy Yates and Tim Cezard discuss the mutability of genomes and the need for versioning in sequence collections.

Nathan Sheffield suggests a Docker endpoint for downloading FASTA files to reduce the load on the ref get sequences service.

Andy Yates emphasizes the importance of pre-populating caches to avoid race conditions and DDoS spikes.

Nathan Sheffield and Andy Yates discuss the need for better communication between tool developers and service providers.

Implementation and Scaling of Sequence Collections

Sveinung Gundersen suggests using Omni Pi for scaling up sequence collections on Kubernetes.

Nathan Sheffield explains the process of creating pedantic objects from FASTA files for database insertion.

Andy Yates and Nathan Sheffield discuss the potential for ensemble to adopt the current implementation for ref get sequence collections.

Nathan Sheffield offers to help with documentation and implementation of sequence collections.

The team agrees to explore the use of sequence collections for CRAM files and improve discoverability through APIs.

Next Steps and Meeting Schedule

Nathan Sheffield raises issue 88, which involves finding collections that contain specific sequences.

The team discusses the potential for adding this functionality to the API or keeping it as an optional feature.

Reggan notes that the next meeting is scheduled for April 30, but they decide to cancel the May 14 meeting.

The team plans to organize a meeting with CRAM people to discuss the proposed changes and improvements.

The meeting concludes with a plan to continue working on the discussed topics and schedule future meetings.

2025-03-19: Schema

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:Sveinung , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Add content to the slide deck for connect	Andy, Tim,Nathan,Sveinung	26-March ?
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Update on comms. - Press release date - 27th March	
4.	Discussion topic: Finalizing the schema; is the one in the spec an example, and where should the final schema live? https://ga4gh.github.io/refget/seqcol/	

5.	Any updates on connect slides prep, reminder of connect agenda	
6.	Demo of new comparison interpretation tool: https://7ca96047.refget.pages.dev/comparison	
	A.O.B. 1) **Website content for refget Sequence Collections** - Website content for refget Sequence Collections **2) Connect session - Wednesday, 2 April 2025, 16:00 - 17:30 ET** Agenda template here - W GA4GH Connect April 2025 - Refget Sequence Collections Standards .docx Presentation template here - P GA4GH Connect 2025 - Refget Sequence Collections Standards .pptx	

Meeting recording:
https://us02web.zoom.us/rec/share/oeHKQ2ZI9NWIim-2OmvwilWAOvRigCRbBw4sKwzMXtLrjN4h0rr3--wZieqpH0Hm.OvMP-uEf7TSEu_EZ
Meeting minutes:

Summary
The team discussed the finalization of the sequence collection schema, agreeing to include attributes like sorted name-length pairs. They proposed versioning the schema semantically and hosting it in a dedicated GitHub repository. Nathan demonstrated a comparison interpretation tool, suggesting it could be extended to an API. They reviewed previous actions, including updates to the paper and a press release with a new date of March 27. The team also planned for the Connect session, with Nathan and Andy preparing slides on the approved standard and future directions. They emphasized the need for clear documentation and organization in the GitHub repository.

Discussion on Algorithm and Implementation Details

Tim Cezard and Nathan Sheffield discuss the simplicity of the algorithm and the use of a library instead of the RFC.

Nathan Sheffield admits he didn't thoroughly check his implementation, while Tim Cezard mentions they didn't verify theirs either.

Nathan Sheffield plans to show a new tool he's working on for comparison interpretation.

Review of Previous Actions and Paper Updates

Nathan Sheffield reviews the agenda, including previous actions, updates on the schema, and Connect slides preparation.

Tim Cezard mentions he reviewed the paper and suggests a detailed description of each implementation in the supplementary.

Nathan Sheffield agrees and suggests having a use cases section and an implementation section in the paper.

Discussion on the clarity of the use cases and implementation sections, with suggestions for concrete examples and real analyses.

Timeline and Paper Readiness

Nathan Sheffield inquires about the timeline for completing the paper, with Tim Cezard suggesting summer as a more realistic target.

Tim Cezard shares an example of a supplementary from the red cat paper to illustrate what they could include.

Nathan Sheffield and Tim Cezard discuss the readiness of the paper and the need for a timeline.

Reggan updates on the comms team's progress and the new date for the press release.

Comparison Interpretation Tool Demonstration

Nathan Sheffield demonstrates a comparison interpretation tool he's developing, explaining its functionality and potential uses.

Discussion on the complexity of interpretation and the need for different modules based on the problem being addressed.

Tim Cezard and Speaker 5 provide feedback on the tool, suggesting additional modules and the potential for an API.

Nathan Sheffield considers making the tool an API and adding different boxes for coordinate system compatibility, sequence comparison, and subset analysis.

Connect Slides Preparation and Finalizing the Schema

Nathan Sheffield reviews the Connect agenda and the need for slides on the approved standard and implementations.

Discussion on the final schema, including whether it should be included in the spec or as an example, and where the final schema should live.

Nathan Sheffield proposes using the schema registry for long-term storage and hosting the schema in the GitHub repo for short-term use.

Tim Cezard and Speaker 5 discuss the organization of the GitHub repository and the need for clear links to the specifications.

Schema Versioning and Organizational Structure

Nathan Sheffield explains the concept of semantic versioning for the schema and proposes setting up the schema for this approach.

Discussion on the need for an input schema and an output schema, with Tim Cezard suggesting a minimal schema first and adding attributes later.

Speaker 5 proposes splitting the schema into a vocabulary and a minimal schema, with references to the full properties file.

Nathan Sheffield agrees to implement the proposed changes and update the example in the spec to point to the final schema.

Final Remarks and Next Steps

Nathan Sheffield summarizes the action items, including creating a JSON file for the schema, updating the example in the spec, and organizing the GitHub repository.

Reggan confirms the completion of the slide deck by the 26th of March.

Nathan Sheffield reminds everyone of the Connect session and the need for slides on the roadmap and ensemble use case.

The meeting concludes with a reminder to complete the slides and prepare for the Connect session.

2025-03-05: Compliance test and service info

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Reggan Thomas ,Sveinung Gundersen , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Review and add to paper.	All	
2	Reggan to try to get comms draft by 18 March	Reggan	Done
3	compute top-level digests for all demo_fasta files https://github.com/refgenie/refget/tree/master/test_fasta let's compare them to make sure we're doing it the same way.	Sveinung, Tim	2 weeks. Done
4	Tim to update tasc artifacts to 'refget-sequence' and 'refget-seqcol'	Tim	Done
5	reach out to pangenome session to see if they would like a short presentation on "unique identifiers for pangenomes"	Reggan	Done

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	

3.	Discuss comms team draft - In progress draft will be available by 20-March	
4.	Compliance tests	
5.	service info artifacts: https://github.com/ga4gh/refget/pull/93	
6	Connect prep Connect session - refget Sequence Collections standard Date & Time **Wednesday, 2 April 2025, 16:00 - 17:30** Agenda template here - W GA4GH Connect April 2025 - Refget Sequence Collections Standards .docx Presentation template here - P GA4GH Connect 2025 - Refget Sequence Collections Standards .pptx	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/nFG_7ShsQf-7tmJYA-dpmN-X2k9BWlzDbktGMVmPI19mWC1XCu2CzLIkyPR2HYY-.8mCfrgev77WjR1RT
Meeting minutes:

Nathan Sheffield:	https://refgenie.org/refget/notebooks/seqcol-client/
Nathan Sheffield:	https://refgenie.org/refget/compliance/
Nathan Sheffield:	https://github.com/refgenie/refget
Nathan Sheffield:	https://github.com/refgenie/refget/blob/master/tests/test_digests.py
Nathan Sheffield:	https://github.com/refgenie/refget/blob/master/tests/test_refget.py
Nathan Sheffield:	https://github.com/refgenie/refget/tree/master/test_fasta

https://docs.google.com/document/d/1JF6Q3NCifbdPwJLToI0cm0oFWPZwmY4e/edit?rtpof=true&sd=true&tab=t.0

Summary:
The team discussed the agenda, focusing on previous actions, compliance testing, and the draft press release. Nathan proposed finalizing the manuscript's focus on the spec and adding use cases. Tim and Sveinung debated the manuscript's technical versus use-case content. Nathan emphasized the need for compliance tests, requesting Tim and Sveinung to compute digests for demo FASTA files. Tim mentioned the service's operational status and the need to update the spec to reflect new artifact naming conventions. The team also planned presentations for the Connect event, including a 15-minute overview by Nathan and five-minute updates on implementations.

Actions:
Compute top-level digests for all demo FASTA files and compare them to ensure consistency.
Update the ref-get package to version 2.01, deprecating the use of "ref-get" and transitioning to "ref-get-sequence".
Reach out to the Pan Genome session chairs to see if they would like a 5-minute update on the ref-get work during their session
Finalize the schema URL for the spec and discuss it at the next meeting.

Discussion on Manuscript Content and Structure
Nathan explains the current manuscript's purpose as a brief advertisement of the spec, focusing on the big picture and how the spec fits into the world.
Sveinung Gundersen and Tim Cezard discuss the need for more technical background and analysis in the manuscript.
Nathan suggests two papers: one on the spec itself and another on the analysis of common reference genomes using the tool.
The team debates whether to include use cases in the spec paper or leave them for a follow-up paper.

Implementation and Documentation of the Spec
Nathan discusses his work on the Python package for the ref get spec, including tutorials and user-friendly documentation.
Sveinung Gundersen suggests including examples and use cases in the documentation to make it more understandable and engaging.
Tim Cezard proposes adding use cases to the paper to illustrate the spec's potential applications
Nathan agrees to include use cases in the paper and suggests focusing on the spec and use cases in the initial paper.

Compliance Testing and Implementation Updates
Nathan raises concerns about the status of his implementation and the need for confidence in compliance tests.
Tim Cezard mentions the need to implement the last two endpoints of the spec and suggests restarting the server to ensure the correct digest computation.

Tim Cezard proposes adding use cases to the paper to illustrate the spec's potential applications.
Nathan proposes creating a bank of demo FASTA files with their official digests for compliance testing.
The team discusses the need for multiple implementations of the digest computation to ensure accuracy

Service Info Artifacts and Naming Conventions
Tim Cezard discusses the naming conventions for artifacts, suggesting singular naming and the use of hyphens
Nathan and Sveinung agree on the use of hyphens for URL-friendliness.
The team discusses the need to coordinate with existing services that use the old naming conventions.
Tim proposes updating the spec to reflect the new naming conventions and possibly deprecating the old names.
Connect Preparation and Presentation Details
Nathan outlines the Connect event agenda, including his 15-minute presentation on the approved standard and five-minute presentations on implementations.
Tim Cezard raises the issue of coordinating with the human pangenome consortium for their session.
Reagan suggests using the comms form or emailing the session chairs to coordinate with the pangenome group.
Nathan proposes giving a five-minute overview of the spec at the pangenome session if invited

Finalizing Schema and Specification Details
The team discusses the need to finalize the schema URL and ensure it is correctly defined in the spec
Tim Cezard suggests defining the schema as text in the spec to avoid confusion about whether the example is part of the spec.
Nathan mentions the upcoming session on the schema registry at the Connect event
The team agrees to finalize the schema and its URL in the next meeting

2025-02-19: Connect proposal

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Vasu (Sanger) ,Sveinung Gundersen , Sasha (GA4GH) ,Reggan (GA4GH) , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

1	Review manuscript https://docs.google.com/document/d/1sp2IhaHsemelA8IaCBwPhZ_uEdN7XGyWQWPdL5w7ckU/edit?tab=t.0	All who want to contribute	2 weeks
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Review issues raised on spec during final review	Done
4.	Review outreach coordination	Done
5.	Review the priorities for this year.	Done
	A.O.B. Connect session - refget Sequence Collections standard Date & Time **Wednesday, 2 April 2025, 16:00 - 17:30** Agenda template here - W GA4GH Connect April 2025 - Refget Sequence Collections Standards .docx Presentation template here - P GA4GH Connect 2025 - Refget Sequence Collections Standards .pptx Deadline - **Feb 26,2025** **Future Opportunities for the Product:** 1. **Webinar:** The comms team can help set up a webinar for product education, where product leads can explain the background, implementation, and other key aspects. 2. **Technical Blog Post:** The product team can write an in-depth piece about the specification, and the comms team can assist in publishing it on the website**.**	

Pangenome:

Connect sessions **:**

1) **Pangenome structural variant reporting and representation**
Friday, 4 April 2025, 11:15 - 12:45 ET

2) **Identifying core consent elements for pangenomic studies**
Friday, 4 April 2025, 14:15 - 15:45

Qatar Genomics:

- **Qatar genomics contact - Hmbarek@qf.org.qa (Hamdi Mbarek)**
- **Radja Badji** - rbadji@qf.org.qa (DP1)

Human Pangenome project:

Community Meeting on March 4th, 8am PST/11am EST/4pm GMT
Agenda
https://ga4gh.slack.com/files/U02MPFRB0G3/F08DNUY09AS/agenda-march2025-hpp.pdf

Chair: Karen Miga

Meeting recording :
https://us02web.zoom.us/rec/share/wXOFRwNKL7OpHnp1DTogBWPqrAt3qlJH4wr-ncwumzPORxFAS5_Z5sxg7L907uew.o11YJNX55OaZqndN
Meeting minutes:

Previous action items

1) Comms team to announce approval, get anything needed from us

- RT - The comms team met with Tim on February 13 and gathered the required information for the press release.
- **Next step**: Jaclyn (Comms) will draft the press release document..

2) Update refget landing page to reference the 2 refget standards

https://www.ga4gh.org/product/refget/

- RT - This is part of the communications request. An overarching page will be created, with links to subsequent pages for additional products.
-
- Publishing Press release
- Creating product webpage
- Promoting across our channels

-

3) Reach out to Rob, Shakuntula, and other contributors to see if they want to coordinate institutional press releases.

- RT - I sent the email on February 7 and received a response from Alex, who suggested highlighting the alignment with VRS 2.0 and vice versa.
- Other request received are:
- - elaine.harrison@elixir-europe.org
- Ingeborg.Winge@uib.no
- Share this with GA4GH Comms team

GIF:
Open call - P GIF-Project-Open-Call-advert.pptx

I'd like to share an exciting opportunity with you all—The GA4GH Implementation Forum (GIF) has opened a new call for **GIF Projects**!

GIF Projects are **community-led initiatives** aimed at advancing genomic data interoperability by implementing GA4GH products. These projects focus on **practical solutions** to improve data discoverability, sharing, accessibility, and integration across the genomics ecosystem.

Projects with strong potential will have the chance to **present at the 2025 April Connect meeting**, providing a platform to showcase their work, gain visibility, and engage with the broader community.

If you have an idea, **we encourage you to propose a GIF Project!** You can find more details and guidelines at https://www.ga4gh.org/open_call/propose-a-gif-project/

NS - https://ga4gh.github.io/refget/
NS - https://github.com/ga4gh/refget/pull/91

NS - https://github.com/ga4gh/refget/issues/79

Summary:
The meeting focused on finalizing the sequence collection standard for press release, with key tasks including reviewing previous actions, updating the splash page, and preparing for the Connect session on April 2. The team discussed the inclusion of the accession attribute in the schema, deciding to leave it out for now due to inconsistencies. They also planned to decommission the HDS spec version and redirect to the new web page. Future directions include integrating with tools like BED and CRAM, and developing services for pan genomes. The team also considered promoting the standard through webinars and technical blog posts, with a draft publication needing feedback.

2025-02-05: Decommission hts-spec version and edit spec

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Comms team to announce approval, get anything needed from us.	Reggan	2 weeks (In progress)
2	Read through and edit the spec itself one last time: https://github.com/ga4gh/refget/blob/dev/docs/seqcol.md - [x] Update to "approved" - [x] Update to "v1.0.0"	All!	2 weeks In progress
3	Decommission hts-spec version of refget sequences documentation, redirect to new web page. - fix link here: https://www.ga4gh.org/product/refget/ - add notice to, or remove, old markdown document, here: https://github.com/samtools/hts-specs/blob/master/refget.md https://github.com/andrewyatz/hts-specs/blob/feature/specification_move/refget.md update ready - review/port over any relevant issues on the hts-spec repo: https://github.com/samtools/hts-specs/issues?q=is%3Aissue%20state%3Aopen%20refget Of the open issues, https://github.com/samtools/hts-specs/issues/771 was the only one that persists as an open issue. Molecule was another one but seems better to re-evaluate in the light of seqcol now being here	Andy/Tim	2 weeks Done
4	Splash page updates: - [x] add ga4gh logo - [x] add link to use cases section, from the splash page	Nathan	2 weeks
5.	Update refget landing page to reference the 2 refget standards https://www.ga4gh.org/product/refget/ - change this to "refget sequences" - create new page for v1.0.0 of refget sequence collections - interlink them, so they point to one another.	Reggan	2 weeks In progress
6.	Reach out to Rob, Shakuntula, and other contributors to see if they want to coordinate institutional press releases.	Reggan	2 weeks

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Update on steering committee, connect proposal. Product is approved, next thing is Reggan will work with comms team. about 2 weeks for news to be sent out. Connect proposal: still working, unofficially, April 4th 9-10:30.	
4.	Go through Issues github issues together https://github.com/ga4gh/refget/issues	
5.	Discussion on pangenome/next steps?	
	A.O.B.	

meeting recording:
https://us02web.zoom.us/rec/play/hx8Q-_R9YJviLXkgQK9tD6xzI46PYZ_CRtb6BJ4IOd3n3j6v0r05mrypdgTk4mjmIQk0DiJfMEwkr93C.ty8nQUv1tzQm9_bj?accessLevel=meeting&canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2Fi5QYwGEqHx0Qk6iHM_iHYzvSQIJXjsvXVrjoCWXsUiUTrnYEsKki_VoKXnnfP64I.0dK3ESUY1C0vn-kY

Meeting minutes:

Summary:
The team discussed the approval of the ref get sequences standard by the steering committee and the next steps for communications, including a tentative timeline of 10-15 days for the announcement. Reagan will work with the comms team to finalize content. Andy was tasked with decommissioning the old ref get documentation and updating links. The team also planned to polish the specification and event page before the announcement. They decided to transition the meetings to focus on pan genomes starting March 19, with the final sequence collections meeting on March 5. The Connect proposal is scheduled for April 4, and the team will coordinate press releases with various institutions.

Next Steps for Communications and Confirmation

- Reggan discusses the timeline for communications, including filling up a comms form and waiting 10 to 15 days for the comms team to act.

- The need for content for the comms team is highlighted, and Reagan plans to talk to the comms team to determine the type of content needed.

- The importance of reviewing the content before the next meeting in two weeks is emphasized.

- The discussion includes the need for a final review of the specification and the event page.

Web Page and Specification Updates

- The need to update the web page and the specification is discussed, including the landing page and the use cases section.

- The current state of the web page is reviewed, and the need to finalize the transfer of information between the old and new sites is mentioned.

- Andy is assigned the task of decommissioning the old HTS spec version of the ref get sequences documentation and redirecting to the new web page.

- The discussion includes the need to update the specification to version 1.0 and make it official.

Open Issues and Task Assignments

- The current open issues on the HGS specs repo are reviewed, with a focus on those tagged with "ref get."

- Andy is assigned to review and possibly close or transfer old issues to the new repo.

- The need for a final read-through of the specification by all participants before the comms announcement is emphasized.

- The importance of coordinating with the comms team and other institutions for joint communications is discussed.

Coordination with Other Institutions and Timelines

- The need to coordinate with other institutions like EBI, UVA, and Alex here for joint communications is highlighted.

- The importance of aligning timelines for announcements and ensuring all institutions are on the same page is discussed.

- The need to reach out to other contributors to ref get and coordinate institutional press releases is mentioned.

- The discussion includes the need to update the landing page on the GA4 GH site to reflect the new standard.

Pan Genome and Sequence Collection Meetings

- The discussion shifts to the future focus of the meetings, with a focus on pan genomes and sequence collections.

- The idea of splitting the meetings into separate pan genome and sequence collection meetings is considered.

- The need to balance the focus of the meetings to ensure both topics are addressed effectively is discussed.

- The decision is made to continue the current meeting format and transition to pan genome meetings in the next few weeks.

Connect Proposal and Upcoming Events

- The Connect proposal is discussed, with a rough gate scheduled for April 4.

- The need to finalize the proposal and prepare for the Connect meeting is emphasized.

- The importance of having services available for the Connect meeting is discussed, with a focus on demonstrating current services and plans for future services.

- The need to coordinate with other institutions and projects for the Connect meeting is highlighted.

Engaging with Other Projects and Communities

- The need to engage with other projects like HBP, Qatar Genome Project, and non-human pan genomes is discussed.

- The importance of ensuring the standard is applicable to a wide range of genomes and projects is emphasized.

- The discussion includes the need to reach out to potential collaborators and ensure they are aware of the standard and its development.

- The importance of aligning the standard with the needs of different communities and projects is highlighted.

Final Preparations and Action Items

- The need to finalize the web page, specification, and other materials before the comms announcement is emphasized.
- The importance of coordinating with the comms team and other institutions for joint communications is reiterated.
- The discussion includes the need to update the landing page on the GA4GH site and ensure all materials are ready for the Connect meeting.

2025-01-22: PSC report / Interop

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Report on steering committee meeting.	
4.	Interoperability, and plans for the EVA service updates. Basically, discussion on plans for deploying instances	
5.	Go through Issues github issues together https://github.com/ga4gh/refget/issues	
	A.O.B.	

meeting recording:

meeting minutes:

2025-01-08: PSC

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	alias to collection : is this good enough ?	Tim?
4.	Comments from Mike Love?	MIke
5.	Confirm/finalize slides for Jan 13 presentation. Steering Committee presentation Slides	Nathan
6.	Need to complete the following sections of **this document** (product approval submission form): 10.1.3 - **Potential adopters** 10.1.4 - **Problem statement** 10.1.5 - summary of **adopter view** 14.8.3 - **implementations - demonstrate interoperability)**	
7.	Go through Issues github issues together https://github.com/ga4gh/refget/issues	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/S7RKrbZsbNX3wXii6xnTZ1SvLv26tqt2NdB7riv0i1IHb8wh82Ne5T8vGG3cp8kQ.5bz71W-XOxdcwD0N

meeting minutes:

The meeting focused on finalizing the presentation for the January 13th review of the ref get sequence collections standard. Key points included confirming the presentation time (9:15 AM BST) and location, and ensuring the slides are ready. The team discussed the need to demonstrate interoperability and provide a timeline for updates. They highlighted potential adopters like Ensembl, ENA, and NCBI, and emphasized the importance of aligning with the VRS standard. Future directions include ref get pan genomes and collaborating with HPRC. The team also discussed the need to update the product approval submission form and ensure all links are current.
Action items:
Contact Tim to get an update on the timeline for his implementation to be up-to-date with the latest changes.
Demonstrate interoperability between Nathan's implementation and a client.
Finalize the test suite and be able to show it working on Nathan's service.
Add highlights of 2-3 key PRC comments and how the project responded to them in the presentation slides
Add the contributor list to the main ref-get documentation.
Check if the presentation time can be extended from 10-15 minutes to 15-20 minutes.

Agenda and Presentation Preparation
Nathan Sheffield outlines the agenda, focusing on confirming and finalizing slides for the January 13th presentation and completing the product approval submission form.
Andy Yates discusses the expansion of points flagged by Reagan, including potential adopters and problem statements.
Nathan Sheffield and Andy Yates debate the interoperability of implementations and the need for a test suite to demonstrate compatibility.
They discuss the challenges of demonstrating interoperability due to recent changes and the need for a timeline for achieving compatibility.

Interoperability and Compatibility Concerns
Nathan Sheffield and Andy Yates discuss the interoperability of different implementations and the need for a test suite to demonstrate compatibility.
They consider the possibility of demonstrating interoperability with a remote service and a local client.
Nathan Sheffield proposes contacting Tim to get a timeline for updating the implementation to the latest changes.
Andy Yates suggests stating a link to the test suite and providing dates of the last test runs to demonstrate compatibility.

Potential Adopters and Use Cases
Andy Yates lists potential adopters, including Ensembl, ENA, NCBI, and other NSDC repositories.
Andy Yates explains Ensembl's new infrastructure and the potential for providing sequence collections through a central metadata database.
They discuss the potential for Ensembl and other projects to provide public services for sequence collections.
They discuss the challenges of mutable genome assemblies and the need for a versioning system to handle changes.

2024-12-11: Github issues

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": , , , , , , ,
Apologies : Sveinung

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Go through github issues, close or milestone or comment (add issues that need discussion to next mtg agenda)	Tim/Nathan	
2	Prepare slides for Jan 13th steering committee meeting	Tim/Nathan	
3	Give Jan 13th steering committee presentation (15 min)	Nathan	Jan 13
4	Cancel Dec 25th meeting; next meeting will be Jan 8th	Reggan	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Anything final on the spec? Close github issues	
4.	Presentation for steering committee meeting	
	A.O.B. - Connect meeting - 1 to 4 April 2025 • Broad Institute of MIT and Harvard, USA Register for Connect	-

	- Add your connect session proposal to the Easyretro board here by 6th Jan Need to submit proposals in the form by 13th Jan - Next meeting is on 25th,Dec - Cancel? Jan 8th,2025 - First meeting of the year.	

Meeting recording:
https://us02web.zoom.us/rec/share/bnUhwsDzsE-rGnjsbqJx_vcCqcFPYwaijq1N3Ft8ckLoLKLFDDHCueUX5bgAn8_Y.oiGoy1kM3NI3IZ0I

Meeting minutes:

- We submitted a connect session proposal for the April Connect at the Broad.

2024-11-20: Issue 84,86

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Sveinung Gundersen , Vasu ,Reggan Thomas , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Review the spec	Sveinung / Tim	
2	Cancel meeting on 27th Nov	Reggan	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Advanced attribute qualifiers: passthru and transient attributes https://github.com/ga4gh/refget/issues/86	
4.	Alignment: inherent property https://github.com/ga4gh/refget/issues/84	
5.	A.O.B.	

	1) Connect meeting - 1 to 4 April 2025 • Broad Institute of MIT and Harvard, USA Register for Connect Fill out the session proposal form by 5 December 2024. Next meeting is on Dec 11thm,2024	

Meeting recording:
https://us02web.zoom.us/rec/play/P6QNz2PUGjE5YMOEWlfQUijdWEFTMelonJarte2d8BmcbCBzuTCv4EKTbo6spi8VdyYeesYD2HC3Hgv4.tUd_nddoZgdetG8i?canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2FXZNFc0ZkvtSOZe6gq1GkB-aEj8cOb3rK57PKrXWU2A2fVODz1bxIESmkKWW428hA.dEJcXqk3pna68n89

Meeting minutes:
https://github.com/ga4gh/refget/issues/86#issuecomment-248872829

Summary:

The meeting focused on the attributes endpoint, clarifying that pass-through attributes cannot be used in the attribute endpoint and that transient attributes should not be returned in the collection endpoint. The list endpoint should allow filtering using any attribute except pass-through, with a note that pass-through attributes may not necessarily be digests. The group agreed on the importance of maintaining the schema's integrity and discussed the practical implications of these decisions. They also touched on the inherent property issue, deciding to move it under the GA4GH qualifier. The next steps include finalizing the spec, preparing for the steering committee meeting in January, and proposing sessions for the Connect meeting in April.

action items:

- Finalize the spec based on the decisions made in the meeting.
- Review the GitHub issues and close out any remaining items.
- Propose a session for the upcoming Connect meeting.

Discussion on Pass Through and Transient Attributes

- Nathan initiates the meeting and outlines the agenda, focusing on pass-through and transient attributes.
- Nathan and others discuss the confusion between list and attribute endpoints regarding pass-through attributes.
- The group agrees that pass-through attributes should not be used in the attribute endpoint and that the list endpoint should allow filtering using any attribute except pass-through
- The conversation touches on the need for clarity in the specification, particularly regarding the use of digests and string types.

Clarification on List Endpoint and Pass Through Attributes

- Nathan proposes changes to the list endpoint to make it clear that level one representation should be used.

- The group discusses the implications of these changes, including the need for different indexing and searching methods.
- The conversation shifts to the practical aspects of implementing these changes and the potential impact on existing systems.
- The group agrees on the need for clear wording in the specification to avoid confusion.

Handling Transient Attributes in Collection Endpoint

- The discussion moves to transient attributes and their handling in the collection endpoint.
- Nathan and others debate whether transient attributes should be returned in the collection endpoint or omitted.
- The group considers the practical implications of returning transient attributes versus the purity of the data model.
- A compromise is reached, acknowledging the need for transient attributes in certain use cases while maintaining the integrity of the data model.

Constructing Transient Attributes and Comparison Endpoint

- The group discusses the construction of transient attributes and their adherence to the digest algorithm.
- Nathan raises the question of whether transient attributes should follow the same construction process as other attributes.
- The group agrees that transient attributes should follow the same pattern, with exceptions for special cases.
- The conversation touches on the comparison endpoint and the need for equality checks on level one digests.

Finalizing Specification and Next Steps

- Nathan confirms that the specification changes will be made and shared for review.
- The group discusses the next steps, including finalizing the specification and preparing for the steering committee meeting in January.
- The group plans to review GitHub issues and close any that are resolved.
- The meeting concludes with a discussion on the upcoming Connect meeting and the need to propose sessions for it.

2024-10-30: Cancelled (Due to unavailability of the leads)

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B. Three open issues - proceed to obtain PRC approval ? Need confirmation. 1) X Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx	

Meeting recording:
Meeting minutes:

2024-10-16: Retrievability

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Vasu (Sanger), ,Reggan Thomas (GA4GH) Sveinung Gundersen, , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	• Nathan to write up the decisions from the meeting and share a pull request for review. • Team to review the pull request and provide approval to move forward with the first version of the specification.	Nathan	
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Discussion on retrievability	
4.		
	A.O.B. Three open issues to be closed inorder to process for PRC approval 2) Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx	

Meeting recording:
https://us02web.zoom.us/rec/share/mWWCMew0ZduBsBVt0x4mv7LrRSBLwL8Oz5Vnqq3xnjxOzI5Tqh2UAJxE78dbsJ2Z.XS8b9sfQRI32a5to

Meeting minutes:

https://github.com/ga4gh/refget/issues/57

Meeting summary:
The meeting focused on the retrievability of sorted name-length pairs and sequences. Nathan Sheffield proposed two solutions: adding annotations to the schema indicating retrievable attributes or defining sorted sequences and name-length pairs as final digests. Tim Cezard suggested treating these as metadata, while Sunu emphasized the need for standardized sorting procedures. The team debated the complexity of defining both pass-through and retrievable attributes versus just pass-through. They also discussed the urgency of addressing three open PLC issues for the upcoming PRC steering committee meeting on October 29th, with a potential fallback to January 13th if not resolved.
Action Items

- Nathan to look through the open PRC issues and try to finalize them to be presented at the October steering committee meeting
- The group to continue the discussion on the "retrievable" vs "pass-through" concept asynchronously and documents it in an issue on GitHub.

Implementation and Proposal for Retrievability

- Nathan Sheffield describes his local implementation of sorted name-length pairs and sequences, highlighting the issue with current digests.
- Nathan proposes two solutions: adding an annotation to the schema indicating retrievable attributes or defining sorted sequences and name-length pairs as final digests.
- Nathan explains the potential downside of defining attributes as final digests, which could lead to re-digestion.
- Nathan seeks feedback from other participants on his proposals

Feedback and Historical Context

- Sveinung recalls a similar question raised by Tim Cezard and mentions an issue related to it
- Tim Cezard confirms the issue and discusses his approach to treating sorted sequences as metadata.

- Nathan and Tim discuss the historical context of the issue, including previous discussions on annotating which attributes are digested.
- Nathan revisits the idea of adding a property to the schema indicating which attributes are digested.

Debate on Schema and Attribute Definitions

- Tim Cezard and Sveinung discuss the implications of defining attributes as retrievable or non-retrievable versus pass-through or non-digested.
- Nathan Sheffield suggests that the standard could define pass-through attributes to avoid re-digestion.
- Sveinung expresses concern about the complexity of having multiple schemas and suggests keeping the current schema but adding modifiers for retrievability and pass-through attributes.
- Tim Cezard and Nathan discuss the potential confusion of having attributes that are both digested and retrievable.

Practical Considerations and Use Cases

- Sveinung emphasizes the importance of having a standard for creating sorted attributes and the potential use cases for retrievable sorted sequences.
- Nathan Sheffield argues that for most use cases, retrievability is not necessary and suggests focusing on pass-through attributes.
- Tim Cezard raises questions about the use of attribute endpoints and comparison functions for retrievable attributes.
- The group discusses the practical implications of defining attributes as pass-through or retrievable and the potential for different attribute names.

Next Steps and PRC Approval

- Reggan brings up the upcoming PRC steering committee meeting and the need to address three open PLC issues.
- Nathan Sheffield suggests that they may not be able to finalize everything in time and proposes going with what they have for the meeting.
- The group agrees to think through the issues and decide on the best approach for the PRC meeting.
- Nathan commits to reviewing the issues and finalizing them if possible, with a backup plan to address them in January.

2024-10-02: Sorted Name length

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Sveining Gundersen ,Tim Cezard , , , , ,
Apologies : Reggan Thomas (GA4GH)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	• Nathan to write up the decisions from the meeting and share a pull request for review. • Team to review the pull request and provide approval to move forward with the first version of the specification.	Nathan	Carry
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Please review: https://github.com/ga4gh/refget/pull/85	
4.	sorted_name_length_pairs return value	
	A.O.B. 3) X Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx	

	SC meeting in October - October 29th at 12 pm BST Updated timeline document - Refget API Seq collection-v1- Product_Approval_Timeline.docx	

Meeting recording:
https://us02web.zoom.us/rec/share/4CJx9vfGm-HW_Hr5uvXek75O8MdcaoW499vBF28KRT1bm7k1AdTxtY2N3VOLP14Y.DTA1zgChD47IGwKe

Meeting minutes:
Nathan Sheffield and Sveinung Gundersen discussed the implementation of a coordinate system retrieval feature. Nathan proposed two main goals: retrieving a coordinate system for a collection identifier and retrieving a coordinate system for a given coordinate system identifier.

They debated the best approach, considering the need for a name-length pair attribute and the implications of storing and retrieving data. Tim Cezard emphasized the importance of a clear specification for attribute endpoints. They concluded that required attributes should be names, lengths, and sequences at the collection level, and inherent at the attribute level. They agreed to finalize the specification for an upcoming steering committee meeting.

Action Items

- Nathan to write up the decisions from the meeting and share a pull request for review.
- Team to review the pull request and provide approval to move forward with the first version of the specification.

identifier.
Implementation Challenges and Solutions:

- Nathan explains his realization that the name length pair attribute has not been defined, leading to a need to store name length pairs again.
- Sveinung suggests using objects for the name length pair attributes to directly support both goals.
- Nathan proposes restructuring the data into a list of objects instead of adding an attribute, which Sveinung agrees is a special caseOutline

Coordinate System Retrieval Discussion

- Nathan Sheffield discusses his progress on implementing the coordinate system retrieval feature, mentioning he has been working on it in the morning.
- Sveinung Gundersen inquires about the submission process, and Nathan confirms it has been submitted.
- Nathan and Sveinung discuss the meeting schedule, noting that Tim will join late and they should start without him.
- Nathan outlines the two main goals for the coordinate system retrieval: retrieving the coordinate system for a collection identifier and retrieving the coordinate system for a given coordinate system .
- Nathan and Sveinung discuss the implications of storing name length pairs and the potential duplication of data.

Attribute Endpoint and Sorting Mechanisms

- Nathan and Sveinung discuss the attribute endpoint and the need to sort name length pairs.
- Sveinung explains the concept of sorted sequences and how it relates to sorted name length pairs.
- Nathan and Sveinung debate the necessity of sorting and whether it should be part of the attribute endpoint.
- Tim Cezard joins the conversation, emphasizing the importance of a clear recipe for rendering output and avoiding special cases.
- system for a collection and retrieving the coordinate system for a given identifier.
- Nathan and Sveinung discuss the potential need for a separate attribute for name length pairs and sorted name length pairs.
- Tim suggests defining attributes in the service info to indicate whether they are lookupable Standardization and Interoperability
- Tim suggests defining a uniform return value for the attribute endpoint to ensure interoperability.
- Nathan and Sveinung discuss the complexity of defining different return values for different attributes.
- Tim proposes a level three implementation that could resolve attributes to their original JSON objects, but acknowledges the complexity.
- Nathan and Sveinung agree on the need for a clear definition of attributes and their return values.

Finalizing the Specification

- Nathan summarizes the two main goals for the coordinate system retrieval: retrieving the coordinate or not.
- Nathan emphasizes the importance of finalizing the specification and moving forward with a first version, despite ongoing improvements.

Next Steps and Meeting Conclusion

- Nathan mentions the upcoming October steering committee meeting and the need to finalize the specification for review.

- Sveinung inquires about the deadline, and Nathan notes the importance of getting a first version out to start using it.
- Nathan and Sveinung agree to continue refining the specification and aim for a finalized version for the October meeting.
- The meeting concludes with a plan to review and finalize the specification and prepare for the next meeting in two weeks.

2024-09-04: PR81 & NIH grant

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Reggan Thomas (GA4GH), Sveinung Gundersen, , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Roadmap submission for NIH grant - Update this link -Deadline 26 Aug,2024 - Objectives: 1. pangenome; 2. implementation adoption into file formats		
2	Next time: sorted_name_length_pairs return value		

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	

3.	Discussion points on PR 81:https://github.com/ga4gh/refget/pull/81
	- Should we switch to query params to allow multi-filtering? From: /list/:object_type/:attribute/:attribute_digest To: /list/:object_type/?:attribute=:attribute_digest
	  - yes, switch to query params
	  - define multi-params as AND additive filters
	- Should the `/list/:object_type` be plural or singular: `/list/collection` or `/list/collections`
	  - currently, “attribute” object-type is singular, but “list” object-type is plural.
	  - - [x] Need to define ‘object_type” better in the spec. (make sure you don’t accidentally write collection instead of collections).
	  - switch to singular, so ‘object_type’ is always ‘collection’. so it becomes /list/collection.
	- Should we require all the endpoints? Has this been discussed? For instance, if you are only interested in using the `attribute` endpoint, should `collection` then be required?
		-

	- Decision: we can still leave collection as required.	
4.	Roadmap submission for NIH grant - Update this link -Deadline 26 Aug,2024 - Objectives: 1. pangenome; 2. implementation adoption into file formats	Leads
5.	Alignment on inherent: https://github.com/ga4gh/refget/issues/84	Alex?
6.	Finalize approval of final schema	
7.	sorted_name_length_pairs return value	
	A.O.B. 4) Cancel the meeting on 18th Sep (Plenary week)?Next meeting is on 2nd October. 5) Update Column E - Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx SC meeting in October - Dates not yet confirmed 6) Planning to present Refget Sequence collection poster at plenary - Link here	Reggan

Meeting recording:
https://us02web.zoom.us/rec/share/-nEeK-IW6tPDrwLFr89A0WQj_cT7u2k-tW0VUGn6jT3GFXDNvQWeS4T-V5Frti3R.xh31qNtFAwgTvJJt

Meeting minutes:

NS - Discussion about PR 81

NS - collection end point is required and not as recommended

SG - defining what the partial implementation actually is, because, if you implement but a bit differently then it's not interferable right to them.

Moving to the second point - Plural or singular?

NS -
https://github.com/ga4gh/refget/pull/81/files#diff-eec4820598f36755aa23db8de7b5a04ccd34328d0ddb1c54387e68d599e25468R475-R476

NS - Object type is plural

NS - the slash collection endpoint will return a collection.

NS - Slash collections, endpoints, a list of collections. So the plurality determines what you're getting back

Line 476 - make it more formal

Line 483 - possibly put a different collection.

Option 1- as it is. Deal with the fact that object is in some places singular, and in some places.Plural. So that's feels suboptimal.

Option 2 - /list types -I n the list endpoint the name of the variable, so that we use list type.So it's list, slash list, type and list type can be collections.

Option 3 - We just make the list endpoint. Take object type and have it be singular.

NS - Finalised to option 3 basd on the group feedback
NS - Multi filtering
All in favour of the query parameter.
TC- what of the server wants to allow a single parameter. Only one parameter for filtering?
NS - not completely compliant. we allow to filter one attribute.
NS - any argument for path parameters.
TC- Its simple and clear. While query parameter is multiple. Different ways of using it.
SG - are we allowing multi filtering as part of the specification?
TC- unless specified to be allowed. The query parameter is semantically allowed
SG - how complicated is multi filtering?
NS - Based on how we define.
TC - Would like to keep the word digest. Searching and filtering by non -indexing would be expensive.
Refer - https://docs.google.com/document/d/1ZftAYIKE7zMu8pAAr-oDl3uxUZjSJNtAKyEKDWN2Xho/edit

2024-08-21: PR 81 and NIH Grant

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Roadmap submission for NIH grant - Update this link -Deadline Aug 26, 2024	
4.	Merge PR https://github.com/ga4gh/refget/pull/81	
5.	Finalize approval of final schema	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/play/4NHdbubgYiMTkWTFsgeOrsJe4WznodBeJP47QjPZc9q-ep4Cl6fpX0AM-m5sPRVI7GqhZYBDziuXldfd.SgBX7QftB75EtJ5o?canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2Fa4IZeH5aaKiPv7q3HewnN01N4i9eXdMuAnQh6qIUgDxJoqhhXN8yf6C28y1w6bs.MiQnFLBJBCPNzf1L

Meeting minutes:
https://refget.databio.org/attribute/sorted_name_length_pairs/zjM1Ie9m0zFbqsAnZ6jAJSXuFpKTr40J

https://seqcolapi.databio.org/attribute/collection/sorted_name_length_pairs/zjM1Ie9m0zFbqsAnZ6jAJSXuFpKTr40J

2024-08-07: Functions proposed

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Reggan Thomas (GA4GH) ,Sveinung Gundersen , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Review PR 81: https://github.com/ga4gh/refget/pull/81	Everyone	Will be merged in 2 weeks
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	These are 3 new functions proposed to be integrated into the current spec: - Attribute endpoint: https://github.com/ga4gh/refget/issues/80 - List endpoint: https://github.com/ga4gh/refget/issues/61 - Search function: https://github.com/ga4gh/refget/issues/28	
4.	Defining the schema now (with these new endpoints, should we change the default schema so that lengths is no longer inherent?)	
	A.O.B. 1) GA4GH Progress Report Newsletter - Update this link - Deadline 2nd	

	August 2) Roadmap submission for NIH grant - Update this link -Deadline Aug 26, 2024 3) Update Column E - Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx		

Meeting recording:
https://us02web.zoom.us/rec/share/E3RUT0BvaEHIeonSVQ_8YV2GTx6wj-pXHcwd_TEjolw4UoeYObiSnUQleMxD3e9t.VvFaECEbGmhfZz_w

Meeting minutes:
TC - 2 different standards. Right now. Refget. Sequences and Ref. Sequence collections have 2 different ones.
Merging will bring some complications. Not certain about pangenomes.
No issues in having /seq collection in the top
NS - doesn't have to be fixed into the spec. /sequence endpoint. / list end point
NS - think from my perspective, we can define the sequence separate from the sequence
NS - https://seqcolapi.databio.org/docs#/
TC - We will have to have the discussion about whether pangomies an extension.
NS - we all agree on the list end point.
NS - Filter - give me a collection which has thai attribute value. attribute/ attribute digest
NS - it's essentially just the same as the list endpoint. The return value would be exactly the same as List endpoint. You're just getting a filtered list. Right be you give it 2 more, 2 additional pieces of information, which is which attribute and which attribute value. Do you want to filter?
Attribute end point - what about removing the word attribute.
NS - Slash collection slash, and then attribute name instead of attribute, slash collection, and then attribute name
TC - clear on the usefulness and the purpose of the least standpoint
TC - highlight the usefulness of the attribute endpoint. But it is less clear to me.
NS - Rasco is saying, like we should have sorted sequences.
NS - by making the attribute endpoint to use the sorted sequences.

NS - https://github.com/ga4gh/refget/issues/76#issuecomment-2138666899
NS - very opposed to adding anything to the digest.
NS -The information about what that belongs in the encompassing environment.
SG - Name space should be structured
NS - GA4GH/secol - top level digest.
SG - was wondering about was the list endpoint, whether we that support listing attributes.
- might be needed for some usecases that we represent

2024-07-24: Refget discussion - Cancelled (Due to unavailability of the leads)

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B. 4) GA4GH Progress Report Newsletter - Update this link - Deadline 2nd August 5) Roadmap submission for NIH grant - Update this link -Deadline Aug 26, 2024 6) Update Column E - Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx	

Meeting recording:
Meeting minutes:

2024-07-10: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Shakun, , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Attribute endpoint: https://github.com/ga4gh/refget/issues/80 List endpoint: https://github.com/ga4gh/refget/issues/61 Example of the implemented endpoints at https://refget.databio.org/	
4.		
	A.O.B 7) Update Column E - X Refget:Sequence Collections - v1.0 - PRC Comments tracker.xlsx 8) GA4GH Progress Report Newsletter - Update this link - Deadline 2nd August 9) Roadmap submission for NIH grant - Update this link -Deadline Aug 26, 2024	Leads

2024-06-26:

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": , , , , , , ,
Apologies: Reggan Thomas (GA4GH), Sveinung Gundersen

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	https://github.com/ga4gh/refget/issues/80 https://refget.pages.dev/	

4.		
	A.O.B. Connect meeting is on 16 and 17th September 2024 Submit Connect proposal here. GA4GH 12th Plenary details are available here	

Meeting recording:

Meeting minutes:

2024-06-12: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review New docs are live here: https://ga4gh.github.io/refget	
3.	The final discussion topic to be solved for the PRC review was this one: Determining the required attributes in the minimal schema: https://github.com/ga4gh/seqcol-spec/issues/72	
4.	Use case - Sequence collection - Discussion	Alex
	A.O.B.	

Meeting recording:

Meeting minutes:

2024-05-29: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Alex Wagner ,Sveinung Gundersen ,Reggan Thomas , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Approve new docs deployment strategy using materiel-mkdocs. Demo: https://ga4gh.github.io/seqcol-spec/	Nathan
4.	Use case - Sequence collections - Discussion	Alex Wagner
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/E37Jb6JXqyHkm4asGIpBMQ5xF5F13XgnGFf_F8i9pOQObP1vsS5UGP7xGfqrGizJ.Q0qDa_GHGfVQ6LI7

Meeting minutes:
AW - https://vrs.ga4gh.org/en/stable/appendices/design_decisions.html
NS - https://github.com/ga4gh/seqcol-spec/pull/75/files
NS - https://docs.google.com/document/d/1RV4X7nKLvdYzE6qR26l8h57LSlDQWZB65EEpmX1AtW4/edit
NS - https://seqcol.readthedocs.io/en/latest/decision_record/
NS - network of beacons - checking for variants
NS - How do you translate a variant? That's defined on one sequence collection into.
AW - variant is described not in terms of a collection, but specific sequence that can reside in one or more collections.

AW - The allele registry but another service like this is like clinvar and there are of course a host of other services hosted by the NCBI in various institutions that translate and like lift over between builds and stuff like that.

AW - So it's not just a matter of what sequences are supported, but what sequence collections are supported for these transcription services.
AW - Should support at least 2 sequence collections right
SG - basically a beacon would have mapped their variance to.
AW - If you're registering like a node on this network, The idea is you want to be able to say like I support these sequence collections and if that node provides translator services you want to see I can translate between these sequence collections.

AW - Individually query a new node and just say, hey, do you support this collection? Rather than have the nodes like, you know, on a regular basis update what collections they currently support.
NS - how do you try a variant from one sequence collection to another?
Proposal - what I wanna propose is a tiered approach.

- 1st tier is you look at the sequence. So you basically say, is this sequence present?
- In both collections. And if the digest matches, then my variant can be easily put. You know, right there, immediately.
- so it's easy. I can just use that identifier
- tier 2 is okay our sequences didn't match but we had a nameless pair that matches.
- we'll translate that variant onto the sequence that matched this name length pair
- What if you don't have a name like a pair that matches?
- --I would propose a tier 3 that you could guess with a sequence that had a length match.
- So as long as the lengths matched, you might have 2 sequence collections with slightly different chromosome ones like from Ensemble and UCSC, for example.
- Their sequences differ, but their likes are the same and position 1,257 is actually referring to the same spot.
- So as long as their likes match, I could make a guess that If I switch to that sequence. They might have that identifier and so I can translate it that way.
- So I'm basically proposing that you 1st see if the sequences match, if they do it's easy then you look for the name length pair to match and if they do then that's pretty reliable too.

AW - new way of thinking for me this idea that a sequence wouldn't match.

- A name. And a shared link. That it'd be appropriate to say it's the same variant provided that the nucleotides at the position you check are the same.
- Why do we assume that?

NS -
AW - challenge here is that like the premise that a variant is described in terms of like a name.

- VRS is built on Refget
- So the idea here is that our sequence is a digest and the state at a position on that sequence is what we're defining as the variant.
-

NS - proposal to you is that variants are actually defined not on sequences but on coordinate systems.
AW - I want to query a server and say I'm looking for a sequence collection with this digest.

- And that digest is calculated from like an order agnostic, I, you know, set of ref get digest, right, that is part of the sequences attribute and I think it's the sequences attribute in sequence collections.

- Is that something where we can send that API request to a sequence collection supporting service and just say like give us this digest.

AW - I'm saying here's a sequence collection digest. And give me and let me know whether or not you have this sequence collection and I want you to compute these digests using the sequences attribute.
SG : Let's just assume it's not a central registry that every node has some information about sequence.

- Basically what you're saying is that instead of just having the top-level digest, you also include sort of the One digests as part of that structure, which is just a small.
- "lengths": "20e95aade8e72d399dbf7f82a9e84ba5cc4047dc8d791d62",
- "names": "834e2529dc6262d1b774e19e502e4074a1227f0eb91b45a9",
- "sequences": "78f45f5aa3b36a2a8fe1eec415258a036b3753f69acf05df"
- }
- {
- "lengths": "20e95aade8e72d399dbf7f82a9e84ba5cc4047dc8d791d62",
- "names": "834e2529dc6262d1b774e19e502e4074a1227f0eb91b45a9",
- "sequences": "78f45f5aa3b36a2a8fe1eec415258a036b3753f69acf05df",
- "sorted_sequences": "89sd79z7d89a7s89df78s..."
- }
-

AW - In the top level digest a bit. I know there are 3 inherent properties. There's the names, the links and the sequences.

- and then the sequences are optional. So if
- you have a sequence collection. And let's say it's the same names and links for sequence collection A and B and then C.

- then sequence question B has also sequences. Did those correspond to 2 different top level digests?

2024-05-15: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Sveinung Gundersen ,Vasu (Sanger) ,Reggan Thomas , , , ,
Apologies: Shakun B

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Update Column E - Refget:Sequence Collections - v1.0 - PRC Comments track…	Nathan	
2	Read rationale document and: 1. provide comments/edits 2. give opinion on where this belongs: in spec, in paper, in blog post, or in trash.	Everyone	1 week
3	Send rationale document to Rasko	Nathan	2 weeks
4	Review PR: https://github.com/ga4gh/seqcol-spec/pull/75/files	Everyone	Done
	Review poster content - All hands Elixir meeting	Nathan/Tim/Andy/Sveinung	Done

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Sequence collection rationale document: https://docs.google.com/document/d/1RV4X7nKLvdYzE6qR26l8h57LSlDQWZB65EEpmX1AtW4/edit	
4.	Collected updates to spec: https://github.com/ga4gh/seqcol-spec/pull/75/files Is the schema part of the spec or not?	
	A.O.B. Review poster content - All hands Elixir meeting	Leads / 18-May

Meeting recording:
https://us02web.zoom.us/rec/play/rrwbm5eCCMeouQjF_HW1N5TSsPhDrrDkKTwpuJ3Oh8hcSAILxjx_TQWx9fVKWEnVhLrV0jfgPG_MMRfx.EISK_vS89Hy0Bz8J?canPlayFromShare=true&fr

om=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2FGvN0rfRw8CORMfR3P0njt3VjeE-KWAG7XqdNAQ8zZ3GcIWsGIuIFEEgPeOuhqaVD.wtaYLWtMS6fhGIt8

Meeting minutes:

NS: Discussion on specification - It currently explains “how to” but it dont explain the “Why” portion.
Review this PR - https://github.com/ga4gh/seqcol-spec/pull/75/files
Once reviewed, we can go ahead and complete the PRC comments tracker.

2024-04-17: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,John Li (Epic) ,Shakun B , , , , ,
Apologies: Sveinung Gundersen (might join a bit)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Update Column E - X Refget:Sequence Collections - v1.0 - PRC Comments track…	Nathan	
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Discussion about whether the schema is part of the standard ● The list of accepted properties and their definitions	
4.		
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/TbSP1bo2Y2qwjxwywzSkB-w-cvh-SeAsyHk5_guKMYmWIwMOE94kIeF40irqzv0U.WkDjmlCyy7vftVQ8

Meeting minutes:
NS - The specification sort of outlines How to follow the specification.But didn't really explain our rationale. For why we're doing it this certain way.

Add use cases 1 and 2
the third use case. Is the coordinate system use case. And this is sort of where I was coming from and we're spending was coming from where we said, well we have different annotations.

So our challenge, I think, is to try to come up with a system that will be able to solve all those use cases in a way that's easy.

And yet maintain interoperability so that I single identifier can be used for many different use cases.
The potential and terror probability that we can unlock is not at the top level and identifier.

TC - the potential and terror probability that we can unlock is not at the top level and identifier.
NS - The third one is the idea of the layered algorithm for digest for those use cases in which it's really too much to compute the comparison function.
NS - given that we have all these different strategies. For enabling this balance. I think actually It makes sense for us to define the required and the inherent attributes in a recommended schema.
TC - I feel like the schema itself needs to be part of the spec because I think the property definition are extremely important and like somebody would cover log and change the definition of what the sequence is
NS - https://ga4gh.github.io/seqcol-spec/spec_seqcol/

2024-04-03: Refget discussion

Chair: Reggan Thomas
Attendees “Name (Affiliation)”: Sveinung Gundersen, Vasu, Shakun B,Serena K
Apologies : Andy Yates / Nathan Sheffield / Tim Cezard

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3		
4		
	A.O.B.	

	1) Use cases and acceptance criteria 2) Landscape analysis here 3) Presentation - for steering committee 4) PRC approval - pending. Doc submission to the secretariat next week 12th. 5)	

Meeting recording:

Meeting minutes:

Meeting ended early due to unavailability of the leads.

2024-03-20: Refget discussion

Chair: ~~Andy Yates~~ / ~~Nathan Sheffield~~ / Tim Cezard
Attendees “Name (Affiliation)”: Sveinung G, Rasko L,Haroune H ,Reggan Thomas (GA4GH) , , ,
,
Apologies : Shakun,Vasu,Nathan,Andy

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Respond to Comment 11 from Rasko in PRC feedback	Tim	
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3	sorted_sequence attribute: https://github.com/ga4gh/seqcol-spec/issues/71	
4	Review of the specification in https://ga4gh.github.io/seqcol-spec/ What is missing from the specification for v1: • The sequence collection schema in a prominent location (not part of the digest algorithm) • The list of accepted properties and their definitions • Clarify the link between communities and the fields that should be included and or inherent. • Ensure the specification is not opinionated about the fields that should be used or made inherent. What else ?	Tim

5.	Endpoint to List/Filter/Search collections #28 #61 #27	Tim
6.	Review feedback from the PRC members	Leads
5	Review feedback from the REWS team	Leads
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/7aOLTtcjeUlqdZ9VaRjhguU1LADbf510zaq4hAvs2lKsnOaB-5zV8b7J_CVm8DpS.2LvlpfNINQejhtiW

Meeting Minutes:

2024-03-06: Refget discussion- Seq col spec

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,David Yuan (EBI) ,Vasu (Sanger) ,John Li (Epic) ,Serena Kher(Univ of New York) ,Shakun B (Univ of Mauritius) ,Haroune H (EBI) ,Sveinung G (Elixir Norway),Reggan Thomas (GA4GH1§)
Apologies : Shakun

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	What is the current status of GA4GH effort to get DOI or persistent URLs to specifications? RT - Current efforts are documented here	Reggan	Done
2	Reach out to Andy: what refget docs need to be transferred	Nathan	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	New github-pages docs site, with refget docs integrated: https://ga4gh.github.io/seqcol-spec/	
4.	Review feedback from the REWS team	Leads

5.	Review feedback from the PRC members - https://github.com/ga4gh/seqcol-spec/issues/71 - https://github.com/ga4gh/seqcol-spec/pull/70	Leads
	A.O.B.	

Meeting recording:https://us02web.zoom.us/rec/share/uLjvnuE1fyeJ7c1V2mfnhh7RouydnNNYvUDAdf9B8A8rWJoGtUBSU_K3I_VO446T._ymj431Sm1GfI4H0

Meeting minutes:
From chat:
NS - https://ga4gh.github.io/seqcol-spec/
NS - https://www.ga4gh.org/product/refget/
TC - Andy had used https://w3id.org/ga4gh/refget which is currently not pointing to the latest version but can be updated.

TC - https://github.com/samtools/hts-specs/tree/master/pub

NS - https://github.com/ga4gh/seqcol-spec/issues/71
NS - https://ga4gh.github.io/seqcol-spec/spec_seqcol/
NS-
https://ga4gh.github.io/seqcol-spec/spec_seqcol/#f1-why-use-an-array-oriented-structure-instead-of-a-sequence-oriented-structure

NS- Revamp of the documentation -
NS - we were using the readthe doc documentation system. Couple of things we wanted to change - 1) we are rebranding the refget - Sequence collections and refget sequences are both under refget name.
Took the original sequence protocol documentation and merged in sequence collections specification
Initially the refget was only relevant to sequences. We have identifiers of sequences and use those identifiers to retrieve the sequences. But it wasn't applicable for collection of sequences which might need for genome.
Sequence collections standard is a separate standard that build on Refget that allows us to deal with collection of sequences.

2) Now this is posted on github pages instead of relying on third party service readthe docs.
It was one of the PRC comments from readthe docs to Github pages to simpy the process of rendering the documentation.
Now this open for feedback
In the original repository there are 4 branches. The master branch that deployed the specification. Currently at the doc side.
b) there is a dev branch adding changes before we deploy
Makedocs branch - has the new theme
Basically the dev branch - switched to makedoc branch and added the github actions. That branch is deploying the pagehub branch
Pages branch hosting the github links

Once everybody is onboarded , merge the makedocs branch into dev
TC- 1) IF we migrate all materials to the new site - are we also migrate the issue that are associated with the refget that are currently hosted on the hts spec?

2) There are official links pointing to refget specifications and tend to point to the original one. Have a couple of them in mind and need to be updated afterwards.

Where should we point ppl to the spec? To human readable version? to the website?
We should rename this repo to “refget” ie GA4GH.githu.io/Refget
This will be the main documentation for all things related to Refget

On those pages, we should have links to the **Github markdown files** that are the source of the pages.
This points to the human readable version.

2024-02-21: Refget discussion - PR55 &64

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Shakuntala B ,Haroune H ,Sveinung G , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	fix encode thing	Nathan	Done
2	make markdown walkthrough of demo comparison results	Nathan	Carry

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Refget branding	
4.	Discussion on schema: https://github.com/ga4gh/seqcol-spec/issues/50	
5.	Approve pending PRs: - https://github.com/ga4gh/seqcol-spec/pull/55 - https://github.com/ga4gh/seqcol-spec/pull/64 (ok to merge) -	
6.	Finalise product approval doc - W Refget API Seq Collection -v1.0-product_approval_submission_form.docx	Leads

	A.O.B. Review the blurb The Refget: Sequence Collections v1.0 specification is now open for public comment! Collections of reference sequences such as reference genome assemblies are fundamental to genomic analysis. To make their analysis reproducible and efficient, we require tools that can identify, store, retrieve, and compare reference genomes, or any collection of reference sequences. The primary goal of the Sequence Collections (seqcol) project is to standardize identifiers for collections of sequences. Seqcol can be used to identify and distribute genomes, transcriptomes, or proteomes -- anything that can be represented as a collection of sequences. Please submit feedback via GitHub issues or via email Github - https://github.com/ga4gh/seqcol-spec/issues Deadline for the public to respond by - March 22,2024	Leads

Meeting recording:
https://us02web.zoom.us/rec/share/2d2TD-MH52eIKBAk7nH8KVEwgG8wmqYX2i9YrAm6C_vK5INKqvlYRta38AC6LdTn.Rr_mu9N3RauHh4z5

Meeting minutes
NS - https://github.com/refgenie/refget/blob/dev/refget/seqcol_client.py
TC : Can we say this is a client ? https://seqcolapi.databio.org/links_demo.html
SG : https://seqcol.readthedocs.io/en/dev/specification/#step-2-apply-rfc-8785-to-canonicalize-the-value-associated-with-each-attribute-individually
NS - https://seqcol.readthedocs.io/en/dev/digest_from_collection/
SG - https://seqcol.readthedocs.io/en/dev/specification/#f5-the-ga4gh-digest-algorithm
NS - https://refgenie.org/refget/

NS - https://github.com/refgenie/seqcolapi/blob/master/tests/test_compliance.py

PR 64 - Merge? - added an action item for this PR

2024-02-07: Refget discussion - TASC 41

Chair: Andy Yates / Nathan Sheffield / Tim Cezard

Attendees “Name (Affiliation)”: Haroune H,Michael Love,Sveinung G,Vasudeva,Shakun B,Reggan Thomas
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Ask Rasko about ENA as a driver project for seq col (not required for this approval, but this will take awhile now with new person taking Rasko’s place)	Tim	Carry
	Confirm support from Genomics England, Australian Genomics - Andy says Oliver on board for Austrailian Genomics - Genomics England still pending	Andy	
	agenda item for next time: refget branding		Done
	Create PR to add artifact type for service info with TASC, like this: https://github.com/ga4gh/TASC/pull/41		
	write a python client that consumes the seqcol api	Sveinung	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	Nathan
3.	Use case from Mike	Mike Love
4.	EVA implementation: ▭ Seqcol implementation	Haroune/Tim
4.	Discussion: Refget API - Seq collection v1.0 - documentation • Refget API Seq Collection -v1.0-product_approval_submission_form • Refget API Seq collection v1.0-REWS_DSWS_questionnaire • Product Approval Process Timeline • Approval tracker - GitHub link	
	-	
6.	Discussion on: https://github.com/ga4gh/seqcol-spec/issues/36 Are we satisfied with this limitation? Conclusion: we note the limitation in the spec.	Nathan

	https://github.com/ga4gh/seqcol-spec/issues/50	
6.	A.O.B Connect meeting (April 21 - 24) at Ascona, Switzerland. Hybrid participation available. - Submit your connect meeting proposal here (deadline - Feb 16,2024)	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline
	Review 2 ADR PRs: - https://github.com/ga4gh/seqcol-spec/pull/63 - https://github.com/ga4gh/seqcol-spec/pull/62	all	Merged
	add Tim/Haroune implementation showcase as agenda item for Jan 24th.		Done
	Ask Rasko about ENA as a driver project for seq col	Tim	
	Confirm support from Genomics England, Australian Genomics	Andy	
	Create PR to add artifact type for service info with TASC, like this: https://github.com/ga4gh/TASC/pull/41		
	Add MAINTAINERS section to contributing	Nathan	Done
	write a python client that consumes the seqcol api	Sveinung	
	Follow-up on compliance suite Conclusion: we should just do something small for now on our own	Andy	Done
	Ask 3 reviewers if they will serve on product review committee	Reggan	Done
	Revisit and comment on: https://github.com/ga4gh/seqcol-spec/issues/36	All	Feb 7th
	Revisit and comment on: https://github.com/ga4gh/seqcol-spec/issues/50	All	Feb 7th

Meeting recording:https://us02web.zoom.us/rec/share/F0qgnZIsCcGseWmwe2VvSRjCpDOR23FpQwAWrNf1Vwfs1PGZYFFIZUw-cdvexfVj.mQC5rP3bXaL29_85

Meeting minutes:

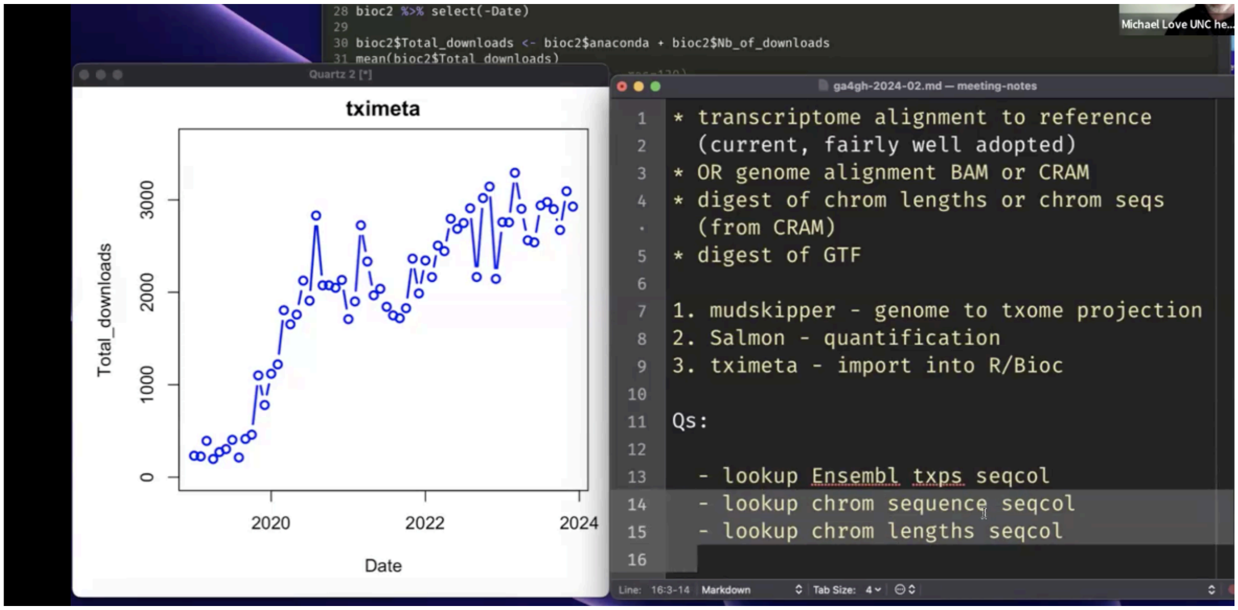
Use case presentation from Micheal Love

* alignment to reference txome (current, fairly well adopted)
* OR genome alignment BAM or CRAM
* digest of chrom lengths or chrom seqs (from CRAM)
* digest of GTF

Qs:

- lookup Ensembl/GENCODE/NCBI txps seqcol
- lookup chrom sequence seqcol
- lookup chrom lengths seqcol

- we have using digest of transcriptome for a while - well adopted pipeline
- Rna seq quantification and they use , user digest of the transcriptome
- wanted to transition to seq col
- Mudskipper (tool) was developed that has projection of genome to transcriptome
- Also looking at genome alignments (BAM or CRAM)
- In future will able to look into ensembl transcript. As a downstream user



```
bioc2 %>% select(-Date)

bioc2$Total_downloads <- bioc2$anaconda + bioc2$Nb_of_downloads
mean(bioc2$Total_downloads)
```

```
* transcriptome alignment to reference
  (current, fairly well adopted)
* OR genome alignment BAM or CRAM
* digest of chrom lengths or chrom seqs
  (from CRAM)
* digest of GTF

1. mudskipper - genome to txome projection
2. Salmon - quantification
3. tximeta - import into R/Bioc

Qs:

  - lookup Ensembl txps seqcol
  - lookup chrom sequence seqcol
  - lookup chrom lengths seqcol
```

TC: Looking for chromosome name and length to calculate seq collection (at the moment you should able to do that right away
or Looking for a way to construct seq collection and look at the server that will give you more info about
ML : the second one (looking for a way to construct)
ML: Imagining API would be helpful like reverse lookup
TC: what we might be missing is a search function.

NS - doesn't need a search function. When there is a server that employs the standard.Currently Tim is working on lookup service and Refgenie will also provide look up service

Seq col Implementation presentation by Haroune

- Scope - Implementation of seq col spec (v 0.1.0)

Seq col ingestion:

- Receive information from trusted source - NCBI
- We cannot accept data from users or other type of sources

Overall design

- Download data from source (NCBI)
- Ingestion end point
- Ga4gh digest

End point

- get service info, get collection, post comparison

NS - Do we need authentication?
HH - Yes
End point demonstration
Future of work
support for additional naming convention
support for fasta file
batch processing

Next steps on Seq collection v1.0
I need to schedule a working session to complete the product approval form and REWS and DS form.

2024-01-10: Refget discussion -

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:
Apologies: Reggan Thomas

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Review 2 ADR PRs: - https://github.com/ga4gh/seqcol-spec/pull/63 - https://github.com/ga4gh/seqcol-spec/pull/62	all	
	add Tim/Haroune implementation showcase as agenda item for Jan 24th.		
	Ask Rasko about ENA as a driver project for seq col	Tim	

	Confirm support from Genomics England, Australian Genomics	Andy	
	Create PR to add artifact type for service info with TASC, like this: https://github.com/ga4gh/TASC/pull/41		
	Add MAINTAINERS section to contributing	Nathan	
	write a python client that consumes the seqcol api	Sveinung	
	Follow-up on compliance suite	Andy	
	Ask 3 reviewers if they will serve on product review committee	Reggan	
	Revisit and comment on: https://github.com/ga4gh/seqcol-spec/issues/36	All	January 24
	Revisit and comment on: https://github.com/ga4gh/seqcol-spec/issues/50	All	Jan 24

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	Nathan
3.	Should we merge concepts of ‘minimal’ and ‘extended’ schemas now? https://github.com/ga4gh/seqcol-spec/issues/50	
4.	Discussion: Refget API - Seq collection v1.0 - documentation • Refget API Seq Collection -v1.0-product_approval_submission_form • Refget API Seq collection v1.0-REWS_DSWS_questionnaire • Product Approval Process Timeline • Approval tracker - GitHub link	
5.	A.O.B Connect meeting (April 21 - 24) at Ascona, Switzerland. Hybrid participation available. - Submit your connect meeting proposal here (deadline - Feb 16,2024) - Previous issue to revisit? Comparison function does not maintain row-wise	

	dependencies when reporting on order	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	
1.	Make a /list end point	Nathan	2 weeks	No longer relevant
2.	Showcase Tim/Haroune's implementation		New Year	

Meeting recording:
https://us02web.zoom.us/rec/play/3UVQebisIU0oqxNJgrYSOGHv2hVhpDHxC4XRl0KyQwdoB_wa-e-d7HG37y2getplGz9pisZ-2dXS7U6d.xkcXv5uxoxAH6MLb?canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2FTFEOL-NB1gP_dVoOQQdUxs8Ju9qSpa9ta0A69EDaMMVT1_fTDC-RKaAfoDgZ076T.BD_IwCe7aJZONbEI

Meeting minutes:
DIscussion on completion of the product approval form

2023-12-13: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": Shakun (uni of Mauritius)
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Make a demo /list end point	Nathan	
2	Showcase Tim/Haroune's implementation		New Year

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	Nathan
3.	Scheduling:	

	- no meeting on December 27th? - next meeting: January 10 - product proposal deadlines (Andy?)	
4.	Minimal and extended schemas proposal #50 https://github.com/ga4gh/seqcol-spec/issues/50 Define what the service info will contain #39 https://github.com/ga4gh/seqcol-spec/issues/39	
5.		
	A.O.B.	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	
1.	Make a /list end point	Nathan	2 weeks	Carry

2023-11-29: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , Sveinung G,Shakun B,Vasu ES,Reggan Thomas, Haroune H
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Dive into the Json schema	all	In 2 weeks
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	Nathan
3.	How to store and represent and compare non collated single value attributes in a	

	sequence collection #57 https://github.com/ga4gh/seqcol-spec/issues/57 See new proposal for compare endpoint array element counts here: https://github.com/ga4gh/seqcol-spec/issues/57#issuecomment-1831979033	
4.	Minimal and extended schemas proposal #50 https://github.com/ga4gh/seqcol-spec/issues/50	
5.		
	A.O.B.	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	
1.	Make a /list end point	Nathan	2 weeks	Carry

Meeting recording:
https://us02web.zoom.us/rec/share/gFZOWTiLJY-tKzPm9sogdG-YMmqBVWeJW07pjAStCu5tC01bFt_p0fXWyCCICI8p.R1T0HRDO8DJFtQP2

Meeting minutes:
Issue 57 - almost done. Noncolated single value . changing the compare end point
Previous action - still working on. action doesn't matter anymore
https://github.com/ga4gh/seqcol-spec/issues/57#issuecomment-1831979033

NS: proposal for tweak to compare endpoint. 3 implementation mentioned in the link.
Named as array elements. Have element A and B. extend to individual attributes
tried flipping around A and B. 2 ways - prioritise the attribute , collection
TC - makes sense. Do we need to split into elements or something?
Spec A contains the number of elements.
NS - A and B counts. Very parallel.
SG: can result in a billion results? It is the same order.
NS : Hierarchical layers.
SG: Just be clear with the single values.
TC: to address the array that are no collated. Anything backward compatible. adding things and not removing things. Compare indivdlual values in a separate property .
SG: Don't have to sort that now? TC: Yeah
NS: Is this solved? Digestion question. Basically not specifying anything.

TC: Start of the issue was single value. but ended up in non collated argument
TC: we are only going to leave to the implementors.
NS: setting up stage. The issue is solved for v1.0
TC: Think the issue should not be closed. Actually close that issue. Unless we want to recreate total
TC: if someone ask about how to specify a single value. we can look at the possibility of how to implement that. Service info matters. How we describe the service info. The implementor decides that.
NS: For now , we are not gonna add any specification for V1.0
Here we should specify how we should be digesting. Spec remains silent on digest on single value.
SG: 3rd one. can switch for changes.
NS: Issue 50
SV: Service info is up to date? Added some comments .
NS : Seq collection has an object. what attributes do you have. ppl could add in their custom attributes. One of them have attribute but other dont. The schema determines tha attribute, in creating the identifier. the service would explain publishing the schema. Should we have a minimal schema.? Idea of extended schema, optional things have to have. Attributes reserved. We provide definitions of the term. The level is at minimal schema and extended. Using schema 2 different way. related issue - 39. how do we actually want to show the schema in service info So we allow the complete schema. How do we define it?
SG: Basically happy about adding the collated inherent to Jason schema. Then no longer compliant with basic schema. The service itself .. Extract collated attributes from Ref.
Jason schema is a combination of subset. Very plain JSON schema. Have them in service info.
TC: We are trying to say what you can find in any seq collection. What can you find in specific implementations? Inherent attribute to be part of the second one. We might want to specify that. As far as JSON schema is concerned. See 2 Json schema. 1st define what can find and 2nd what will be found.
SG: Should we have 2 different schemas?
SB: We want to specify the minimal attributes.? So, if I understand correctly, we want to specify what should be the minimal(inherent) set of attributes that a RefGet object should implement (when working with objects from different providers) along with other attributes that are specific to each provider. Am I correct?
SG: If we dont extend json schema automatically handles.
NS: Json schema is a validation system. but used for defining a structure. The validation is a side benefit. We could use a completely different definition of the object. We are defining structure and well we make it compliant with json schema. It has to be compatible.
SG: its not json schema compliant.
TC: Ref could definitely work. The question about the inherent . Ref would work. Dont see a reason why it wont work
SG : Service endpoint. then we require the implementer to understand. That wont be possible with regular
NS: what we are creating is not a json schema . The validation component. We are proposing 2 things. our schema and json schema are identical which should work.

SV: The jason schema does provide out of box things
NS: Redundancy?
SV: We have 2 list. schema and property list.
NS: Jason with inherent and collated properties.
SG: not under schema.. Then we can create json schema endpoints.
TC: Completely in a different location. behave differently if there is a schematic file. The other defines. Define as a new property. Collated and no collated should have 2 different name. Adding keywords doesn't seem to be a problem. inoperable with different software.
TC: Single json schema for both? we should have 2 different entities.
TC: what should be expected as an output. If you want to add an inherent, we can specify that. Assuming one schema and contains properties to define. using a small subset. Dont expect to find and that should be the purpose of service info.
NS: you can find DEF. you can properties under def. Def as encompassing schema.
SV: Def are small sub jason doc.
NS: Json schema is validation. Validating against the schema, becomes more restricted. set of definition. 2 beyond json schema. One ref to the main schema.
NS: what does the service info provide? Service info presents info in different way.In diff projects we used json schema for years. Definitely allowed to put into the json schema.
NS: challenge - taking out of service info. 2 diff level of hierarchy . the author of the collection or sequence. . Preserve the hierarchy . Need the same structure as json schema. but called it as our model. but basically a json schema. One doc is the subset of other.

2023-11-15: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , Vasu (Sanger),Reggan Thomas (GA4GH) ,Haroune H ,Sveinung Gundersen , , ,
Apologies: Shakun

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

1	Make a /list end point	Nathan	
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review - List endpoint	Nathan
3.	How to store and represent and compare non collated single value attributes in a sequence collection #57 https://github.com/ga4gh/seqcol-spec/issues/57	
4.	Minimal and extended schemas proposal #50 https://github.com/ga4gh/seqcol-spec/issues/50	
5.		
	A.O.B.	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	
1.	Make a /list end point	Nathan	2 weeks	Carry
	-			

Meeting recording:
https://us02web.zoom.us/rec/share/1vbwzSfDJgiTXM3ErZMSFMtZEL1ST98OVQHsz6LIbgqS7R16s1sdzR7maO5q27FU.zOQXXCAVyGoaujG-

Meeting minutes:
Discussion on issue 57

2023-11-01: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,
Apologies: Reggan Thomas

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

1	Make a /list end point	Nathan	2 weeks
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review - List endpoint, pagination -	
3.	Pagination - https://github.com/ga4gh/seqcol-spec/issues/61	
4.	How to store and represent and compare non collated single value attributes in a sequence collection #57 https://github.com/ga4gh/seqcol-spec/issues/57	
5.	Minimal and extended schemas proposal #50 https://github.com/ga4gh/seqcol-spec/issues/50	
	A.O.B.	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	
1.	Review ga4gh pagination https://docs.google.com/document/d/1Zu_ksx6lG3OXcTTViIWyrSMqGTkcpDk9Oa7g1c7StvM/edit	Everyone	2 weeks	
2.	Make a /list end point	Nathan		
3.	Come up with best practices - Pagination: https://stackoverflow.com/questions/55744926/offset-pagination-vs-cursor-pagination https://github.com/ga4gh/TASC/issues/29 - Another idea is to use `?since=id` in combination with a limit, e.g. `?since=d7c638acd72dna73&limit=1000` and have the results ordered by creation date. The user then only needs to keep track of the last id. See e.g. this: https://web.archive.org/web/20180905061056/	Sveinung Gundersen	Done ?	

	https://developer.twitter.com/en/docs/tweets/timelines/guides/working-with-timelines - Compression: https://itnext.io/http-compression-boost-your-servers-speed-a5bad6403452. - User adds `Content-Encoding:gzip` to request compressed result. Server may or may not respond with compressed results. if server can only provide compressed results, it returns a “too large” error unless compression is requested.			

Meeting recording:
https://us02web.zoom.us/rec/play/gPYqomHsYIOjiWQMxcvu4p1ohrzRmt6odekITJ9VsItzva2LkaDa9Z6ajvVCoJgJ3GGjyj6jh5i6mYzL.9P0VLCqbwuSgYzec?canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2FFITwtQiB_gdHnLXdkwNaDFcGSGQAyqPG5o9X4gcInC1O44ZsIhUy9HrGHQZoQu2a.XgjXciI2GjUXO_N7

Meeting minutes:

- previous action: review documentation for pagination, make /list endpoint (issue 61), no implementation yet but in the process of coming up with implementation
- for discussion today: pagination question. Connect feedback: what sequence collection needs for it to be useful. Need list endpoint that provides information on what collections a server holds. Metaserver would be useful for driving adoption but requires each servers make available what they have.
- proposal is to define the API interface and what the return value looks like? Propose there be query parameters, page size and page token. Proposal for return value follows Google pagination. Page size is returning what user put in page token is current pages token and next page token would be the identifier for the next one. There would be a list of collection identifiers
- Items? Should be sequence collections? collection identifiers? propose that it’s a list of strings.
- Page token. Should remain purview of the implementation (encoded offset built into it). Implementers and consumers should assume there’s no predictability.
- Don’t want specific, would prefer items of all resorts and not collection identifiers or collection digests.
- For name of endpoint we should stay away from /collections
- List would be mandatory endpoint. Might be difficulty for implementation that may have large number.Could be recommended.

    - aggregator has to require consistency across multiple implementations and only way to do is through specification level
    - something similar discussed in VRS, go through list of services for particularly query and say this one is returning something rather than adding antenatal list of all possible identifiers
- discoverability of IDs is difficult, moving towards data connect as providing ability to query over time for DRS ids.
- if we took list end point and made it bigger it could be listing a table of the collections with metadata about the collection. Instead of returning list, return a table with metadata.
- Maybe we leave out and move on or address in 2.0
- Data connect and us and DRS should adopt pagination standard that should say how GA4GH things should be paginated. Pagination is built in data connect is toeknization but rather than giving token it says next page. It works the same as proposal.
- Andy will take a look, happy with spec as it stands
- Timothee: using DRS makes assumption of back end implementation
- we will have /list endpoint but return value would look like what their /data endpoint looks like. Data connect says influence table standpoint which we would just return one table ID but still need to provide what's being handed over. Andy will look at how to express it data connect and see if we like it or not.
- /data basically what we propose as return value?
    - if we fully adopt data connect yes, if we want to support alongside list endpoint than no.
- 3 options: list endpoint different from data connect or we borrow their return value, or implement data connect and we have /table and /table id and say our table id is one.
- as maintainers, not specify list endpoint, for implementation we want to expose content of the sequence collection, here is what we propose. We propose you follow data connect standard, se have simplified version where you only expose one table, need to create 3 endpoints but all 3 return static content.
- still have to decide what actual payload would be.
- if we can't implement data connects than there may be something's wrong with data connect. data connect needs to be specialized into one of these use cases? sub standards for different data types?
- Andy will look into it and see if we like it or not
- what we're trying to support in list endpoint is one of the issue data connect tries to solve to be GA4GH then we should be looking at GA4GH standard to follow. other query attributes we would want to have or list and no query attributes?
- what's intersection between JSON? there isn't data connect schema. has table schema.
- In some cases make sense to implement data connect for sequence collection server, does that become part of our spec that you must implement data connect or say sequence collection doesn't provide any list, we recommend using data connect.
- Nathan: would implement endpoint list because it's easy, would not implement data connect because it's overkill. remove from standard and people will not implement listing in general or implement it their own way. And if people started implementing data

connect or it adapts better, few may connect data connect interface. Not done immediately.

- Tim: fine with having it in spec but not making it mandatory. if you want to make it mandatory it should be in v1.
- put something in spec but silent on actual recommendations. work with data connect people for v1.1
    - in agreement with this; don't want to make v1 complicated
    - wait for the right use case to come up.
- Anything you would change about this? items still items? results or records?
    - Timothee: either is fine
    - as long as server can be namespaced. ensure that whatever server was that implemented, they expose at right levels so you could have coexisting APIs that won't clash with list from another endpoint.
- Conclusion: everyone is happy with proposal for /list endpoint. won't be added to spec. in spec we will add blurb about discovery saying the spec is silent on discovery and will be considered for future version. if you need to implement discovery, may consider data connect.

2023-10-18: Refget discussion

Chair: Nathan Sheffield / Tim Cezard/Andy Yates
Attendees "Name (Affiliation)": Sveinung Gundersen,
Apologies : Reggan Thomas (GA4GH)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Review ga4gh pagination https://docs.google.com/document/d/1Zu_ksx6lG3OXcTTViIWyrSMqGTkcpDk9Oa7g1c7StvM/edit	Everyone	2 weeks
	Make a /list end point	Nathan	Carry

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	How to store and represent and compare non collated single value attributes in a sequence collection #57 https://github.com/ga4gh/seqcol-spec/issues/57	
4.	Minimal and extended schemas proposal #50	

	https://github.com/ga4gh/seqcol-spec/issues/50	
5.	List endpoint https://github.com/ga4gh/seqcol-spec/issues/61	

	Previous Actions Arising (label SC/RL/general)	Assigned To	Deadline	Status
1	Make a /list end point	Nathan		Carry
2	Come up with best practices - Pagination: https://stackoverflow.com/questions/55744926/offset-pagination-vs-cursor-pagination https://github.com/ga4gh/TASC/issues/29 - Another idea is to use `?since=id` in combination with a limit, e.g. `?since=d7c638acd72dna73&limit=1000` and have the results ordered by creation date. The user then only needs to keep track of the last id. See e.g. this: https://web.archive.org/web/20180905061056/https://developer.twitter.com/en/docs/tweets/timelines/guides/working-with-timelines - Compression: https://itnext.io/http-compression-boost-your-servers-speed-a5bad6403452. - User adds `Content-Encoding:gzip` to request compressed result. Server may or may not respond with compressed results. if server can only provide compressed results, it returns a “too large” error unless compression is requested.	Sveinung Gundersen		
3				
4				

Meeting recording:
https://us02web.zoom.us/rec/share/z6n491TTgCHpevN_zxMcUhJmbiSqM4p_m1ZCalC5VxyTzebAx9CS-Z4e-c464dyx.bJWclQFBQeAcJvgX

Meeting minutes:

2023-10-10: Reverse Lookup discussion - Cancelled

Chair: Andy Yates / Nathan Sheffield / Tim Cezard

Attendees “Name (Affiliation)”:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			
4			
5			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

Meeting recording:

Meeting minutes:

2023-10-04: Refget discussion

Chair: Nathan Sheffield / Tim Cezard/Andy Yates
Attendees “Name (Affiliation)”: , Shakuntala Baichoo, Haroune H,Andy Yates, Reggan Thomas
, Sveinung Gundersen
Apologies : Rob Davies

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Make a / list end point	Nathan	
	Come up with best practices	Sveinung Gundersen	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	How to store and represent and compare non collated single value attributes in a sequence collection #57 https://github.com/ga4gh/seqcol-spec/issues/57	
4.	Minimal and extended schemas proposal #50 https://github.com/ga4gh/seqcol-spec/issues/50	

Meeting recording:
https://us02web.zoom.us/rec/share/JoOJkozs1PoDehmvxCZxSLX6wHIvmGRrhwUgr3k0uQ9aY7wZoW4-fD66zrbpBOpC.qqdEGq9tQn_k8sWz

Meeting minutes:
NS - each server has an endpoint that returns a bloom filter.so I get back Bloom filters from each. Server and then if I'm given a digest I just compute the the set of agreed upon hash functions.
NS - For that digest from the specified, you know, from the specification of Bloom filter. And then I would query each blue filter and then I would be able to eliminate ones that definitely don't have it
NS - https://broadinstitute.github.io/picard/explain-flags.html
NS - The whole point of this is that We kind of need one place that you can go.Like people sort of want one place that you can go. To learn about your sequence collection. And so it's not that's not really compatible with the different provider
So how do we solve that problem?

NS -what I'm proposing is something like Genbank. Where people can submit sequences should be hosted by somebody like NCBI or Ensemble or one of the big providers.
NS - Do we want to add to the specification like a slash list -Endpoint or something?
AY - not providing that endpoint means that the only way to do that is to ping all possible known servers and wait for one to respond back.

2023-09-12: Reverse Lookup discussion - CANCELLED

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Future of Reverse lookup meetings	Tim
4.		
	A.O.B.	

	Previous Actions Arising (label SC/RL/general)	Assigned To	Deadline	Status
1	Ping readdthedocs previous maintainers for update	Andy		In progress / Completed
2				
3	Compliance document update	Tim		In progress / Completed
4				

Meeting recording:

Meeting minutes:

2023-09-06: Refget discussion

Chair: Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Haroune H ,Shakun B ,Sveinung G ,Rob Davies , , ,
Apologies : Andy Yates

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	switch all trunc512 refs/code to sha512t24u	Nathan	Done
	add: "We sometimes use reference genome, but ..." to emphasize the idea of "Seqcol can be used to identify genomes, transcriptomes, or proteomes -- anything that can be represented as a collection of sequences." or, clarify “reference genomes” -> sequence collections	Sveinung	Done

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Clarifying some review comments: - trunc512 vs sha512t24u	
	Discussion about how to facilitate adoption. ideas: - alpha testers - extremely simple fasta-to-digest tools - comparison between vcf and seqcol. or, BED and seqcol	
	Define service-info: https://github.com/ga4gh/seqcol-spec/issues/39	
	Finalize minimal and extended schemas: https://github.com/ga4gh/seqcol-spec/issues/50	
	A.O.B. Steps towards submitting the product to GA4GH: https://github.com/ga4gh/seqcol-spec/issues/35 Next steps: - two implementations - compliance suite	

2023-08-23: Refget discussion

Chair: Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Haroune H ,Shakun B ,Sveinung G ,Rob Davies , , ,
Apologies : Andy Yates

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	- merge PR 31	Nathan	Done/Done

	- Write an ADR for: no prefix for seqcol UI		
2	have discussion about how to facilitate adoption. ideas: - alpha testers - extremely simple fasta-to-digest tools - comparison between vcf and seqcol. or, BED and seqcol		added to agenda

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Draft of spec available: https://seqcol.readthedocs.io/en/dev/	Nathan
4.	Session description - Please confirm the session description by **Aug 25,2023 -** Presentation template - here Agenda items for the connect session - Agenda template here - Deadline - **05th Sep** Ideas: - start from what we had last time - add sorted_name_length_pairs - showcase schemas - make it more use-case focused (go through demo use cases) - what little bits of tooling do you need to adopt this?	
5.	Define service-info: https://github.com/ga4gh/seqcol-spec/issues/39	Nathan/Tim
6.	Finalize minimal and extended schemas: https://github.com/ga4gh/seqcol-spec/issues/50	Nathan
	A.O.B. Steps towards submitting the product to GA4GH: https://github.com/ga4gh/seqcol-spec/issues/35 Next steps: - two implementations - compliance suite	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline
	Write formal documentation of ga4gh digest: that it's just the algorithm of sha512, trunc, url encode (without prefix). then, prefixes can be added. (PR #31) Revisit PR 31	Andy	Done

	Write an ADR for: no prefix for seqcol UI	Nathan	Carry
	Discuss connect content next time “Sequence collections: A standard for unique identifiers and compatibility of reference genomes” has been approved, and has been scheduled for **Tuesday September 19 at 14:00 - 15:30 PDT**.		
	Cancel the next meeting. we’ll next meet aug 23	Reagan	Done

Meeting recording:

Meeting minutes:
NS - https://stackoverflow.com/questions/52566472/how-to-extend-a-schema-in-json-schema

2023-07-26: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Haroune H,Rob Davies (Sanger), Reagan Thomas (GA4GH)
Apologies: Shakun

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Write formal documentation of ga4gh digest: that it’s just the algorithm of sha512, trunc, url encode (without prefix). then, prefixes can be added. (PR #31) Revisit PR 31	Andy	
	Write an ADR for: no prefix for seqcol UI	Nathan	
	Discuss connect content next time		
	Cancel next meeting. we’ll next meet aug 23	Reagan / Done	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Connect sessions due 30 July 2023	
3.	Continue prefixes discussion: - initial issue: https://github.com/ga4gh/seqcol-spec/issues/37 - ADR for internal identifiers: https://github.com/ga4gh/seqcol-spec/pull/42 - new issue for UI values: https://github.com/ga4gh/seqcol-spec/issues/53	

	- Tim:it’s too early to add/require prefixes	
4.	Which use cases might make sense to include in the specification text? https://docs.google.com/document/d/1aIoA4mhzI8iVnAA1wtDj8GKcp7ATb69i512WLT94F3A/edit#heading=h.vo7w4rw4ya22 See also: https://github.com/ga4gh/seqcol-spec/issues/11	
	Repository of additional fields? Proposals by PR to the “Extended Schema” JSON document?	
	A.O.B. Call for Connect sessions - Submit proposal here. Submission due - **30 July,2023** Call for Poster - Submit proposal here Submission due - **30 July,2023**	

	Previous Actions (label SC/RL/general)	Assigned To	Deadline	Status
1	Review use cases, and pull out or write some for the spec.	Everyone		
2	Agenda for connect session	Everyone		30-July
3				
4				

Meeting recording:
https://us02web.zoom.us/rec/share/UHhBZMThFgERymGCXSjyDC3msfSNgcaZlW1bLhNegihwTsjb2JVL6W0IJCrdmtWN.hNgeXU2vcfUJPbro

Meeting minutes:

2023-07-18: Reverse Lookup discussion

Chair: Andy Yates / Tim Cezard
Attendees “Name (Affiliation)”:
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline	Status
1	Ping readdthedocs previous maintainers for update	Andy		In progress / Completed
2				
3	Compliance document update	Tim		In progress / Completed
4				

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review:	
3.		
4.		
	A.O.B.	

Meeting recording:

Meeting minutes:

2023-07-12: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:Shakun,Haroune,Sveinung G,Reggan Thomas (GA4GH)
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Review use cases, and pull out or write some for the spec.	Everyone	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	

2.	Merge two PRs now: - https://github.com/ga4gh/seqcol-spec/pull/51 - https://github.com/ga4gh/seqcol-spec/pull/49	Nathan
3.	PR 42: Internal identifiers SHOULD NOT be prefixed https://github.com/ga4gh/seqcol-spec/pull/42	Nathan
4.	Any remaining issues that need to be solved before people can start implementing? https://github.com/ga4gh/seqcol-spec/milestone/1	Nathan
5 .	Issue 3: How should we specify the metadata endpoint? https://github.com/ga4gh/seqcol-spec/issues/3#issuecomment-1404011799	Nathan
6.	Which use cases might make sense to include in the specification text? https://docs.google.com/document/d/1aIoA4mhzI8iVnAA1wtDj8GKcp7ATb69i512WLT94F3A/edit#heading=h.vo7w4rw4ya22	Nathan
	A.O.B. Call for Connect sessions - Submit proposal here. Submission due - **30 July,2023** Call for Poster - Submit proposal here Submission due - **30 July,2023**	

Meeting recording:

https://us02web.zoom.us/rec/share/Bvu38RELXv4_nZB8e2dL_bla8ixOcYk1AM3Icvmn83mko1k1LblakZ2DjEY5rdPd.PHAJMnn-tPJWsdcT

Meeting minutes:

PR 42 – Opened in jan 11. We can resolve this. We are not prefixing anything. The sticking thing, what about refget. Digest if we prefix that,
If the refget identifier say, seq call
TC - there's two prefixes. there's two prefixes. and yeah, in refget. We, at least for the GA4GH algorithm, we said that the digest the, the prefix. The optional prefix was a GA4GH codon, but the SQ is mandatory.
SG - Curie identifier. I mean, there was this discussion of the GA4GH party security environments and with that decision, we could define this as a secure identifier.

NS - wouldn't need to include the GA4GH prefix,
NS - for sequences that are included in these things. But if there was some other basically with this, we'd be saying that the sequence attribute expects a GA4GH namespace identifier. and so therefore, we don't have to include it because that's the

default. but if you wanted to put sequences that were being provided by something that was not say refget
NS - ou want to create a sequence collection where let's say, the case matters because your sequences are now sentences say and you need a is C's that a capital or lowercase and so they're not really like refget sequences, but for some other reason, the sequence collection idea, which is now a sentence collection, will be useful for something then sequence you could reuse everything we did

NS – 42 is resolved and requested for review.
NS - we had just decided sequences should not be prefixed with GA4GH but they should be prefixed with SQ dot but then like this comment is basically saying, well, we shouldn't wait SQ dot and I guess I'm I kind of agree with that.

NS - suggesting that we should discontinue with the support for MD five which I think is not going to fly at this point.

NS - I'm going to mark this issue as solved by poor request 42 Because I think it's referring to the construction of the digests that we are then going to have the strings that we're then going to digest for external identifiers, but then I'll probably create another issue that talks about this to just summarise

2023-06-28: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:
Apologies:Reggan

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Two new ADRs: https://github.com/ga4gh/seqcol-spec/pull/49 https://github.com/ga4gh/seqcol-spec/pull/51	
3.	Schema drafts for minimal and extended https://github.com/ga4gh/seqcol-spec/issues/50	

	A.O.B.	

Meeting minutes:

- Nathan created issue with proposed drafts for new ADRs.
- JSON schema schemas are always written in JSON but that's not true. Can be written in YAML.
- LinkML is YAML first but can generate alternate schema representations like JSON schema, python data classes, etc. It's gaining traction. Might be a toolkit idea.
    - Know use case where they went back to JSON schema because LinkML not ready.
- Primary as JSON schema is fine
- Do we allow people to overwrite collated attribute. If you extend, you can overwrite attributes? Guessing some people will want to add.
    - You can add constraint and now object meets it or not. If you extend and add new ref, you are adding additional constraints. Impossible to remove constraint. Can't remove constraint through extension. Only thing you can do is add additional constraints.
    - Do we have a way to say what is or isn't allowed? Articulate schema intends to specify what is allowed or not allowed.
    - What might people want to add to inherent? Typologies? Alphabet?
    - Inherent is definition of array, changing that would change the definition of the array.
- description sentences, take out human readable and add
- Useful to have other names of array, how to map to another array. Take this offline.
- Main concern is whether they are to give informative description of content. There's limited amount of space but these are algorithms and we will have section in spec to describe in detail. Do we need here reference to spec to see if you want more detail to go there.
- adding more detail so that it can stand alone.Could make sure what json schema is referring to follows json LD or not. Keep it in description since it's the authority describing the field. Might make sense to say GA4GH sequence collections specifications and the current version will show up. Document what version this is.
- we should identifier for this schema, not url but curie. question for other schemas also. how to identify the uniquely.
- We want description to provide as much context and authority as possible. Otherwise it needs to refer to other things.
- Add an ID
- Collated is a property? not a keyword. should we define it as a property in our JSON schema and say what should be in it.
- It doesn't really matter because JSON schema processor using this as validator, we're highjacking it as a way to describe it. We're not using to validate objects but maybe we could. Adding collated and inherent not relevant to validation.
- This is outside schema itself, it's at the meta level.
- https://gregsdennis.github.io/Manatee.Json/usage/schema/vocabs.html

- we can write JSON schema that can validate this schema. Service is suppose to extend it and stays correctly. You can define inherent, collated, that's allowed, etc. If someone define their own schema they can use meta schema to validate that.
- Already doing in compliance test for service info, but all done manually and not by schema. might not be necessary but possible.
- push it as genome browser compatibility array and have it as part of description so people have idea of what this is about.
- only thing schema is doing is reserving the name and provide description. Add genome browsers in the description.
- Ok to say it's an intended use case around that feature.
- Stay away from defining it by use.
- Comes down to JSON schema functionality. Metaschema, would that make an argument that we are following how that works.
- We should assume either are possible, assuming no constraints. Need to put in correct place, but collated not in correct place. It's property of property.
- Better as property or leave it outside?
    - better as property, easier for extension.
-

2023-06-20: Reverse Lookup discussion

Chair: Andy Yates / Tim Cezard
Attendees "Name (Affiliation)": Shakun,
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Ping readdthedocs previous maintainers for update	Andy	
2	Check with REWS regarding status of REWS questionnaire	Justina	
3	Compliance document update	Tim	
4			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review: ● Raise with TASC the documentation of existing prefix and process for creating new one: ○ \`SQ\` is pretty much accepted and documented in this thread ○ I've asked for direction on how to chose the new prefixes ● Create a tag of the compliance suite for v1.0 before updates: ○ Tag already exists but need to added to the readthedocs maintainers	
3.	Product approval: PRC meeting #2 on 29th June 2023	

	1. PRC meeting outcomes: Refget V2.0 - PRC- Comments tracker.xlsx 2. Public comments tracker: Refget V2.0 - Comments tracker.xlsx 3. REWS and DSWS: Refget API_V_2.0 REWS-DSWS-issue-tracker.xlsx	
4.	Update [Compliance report]() and [Compliance document]() ongoing	
	A.O.B.	

Meeting recording:

Meeting minutes:

- On prefix we want to use is SQ. Initially described by VRS. Issue on this was raised by VRS on how to register and how to get accepted new prefix. This is established at this point. Left one more comment saying we will use it across 2 GA4GH products.
  - For sequence collection we'll want to use prefix and want a process at that point.
  - There was a comment from Michael, ultimately interpretation was we're doing it.
- Create tag for compliance suite so current state of documentation can be used in readthedocs. Creating the tag is straightforward. To deploy, don't know who has authority to do that?
  - It won't be in the same URL if Tim does it.
  - Ok with going with redeployment but it might be linked to a few places including study documents.
  - Andy will follow up with GA4GH contributor to grant access to space. If we don't get response, we should assume we start fresh in readthedoc space in refget compliance suite. We can always deal with use W3ID URLs and refer to those instead.
  - We have a space at https://w3id.org/ga4gh/refget
  - https://github.com/perma-id/w3id.org/tree/master/ga4gh
  - We can rewrite rule to point to new location and use that. In publications I used this redirect rules more often than direct URLs.
  - https://github.com/perma-id/w3id.org/pull/3361
  - They look to see who should get pinged about the chain
  - Matthew Laird and Sumesh; Andy will ping Matt who should have authorization. Will try to resurrect this first and if we can't, use W3IDs. Put Tim and Andy down as maintainers.
- PRC meeting all done; for actual product approval, we don't have risk assessment completed. Tom Conner from the Broad was going to do it; not done yet, should we ping him? Not a problem for PRC, just needed for product approval/SC meeting.
  - Justina will check with REWS regarding the REWS questionnaire
  - Materials need to be sent to SC on July 3.
  - SC meeting is on July 17.
- Start thinking about SC presentation and we can use some slides we are already have and talk about any changes. There was a presentation used for PRC, might be a good enough state to use with some updates and address what changes we made.

2023-06-14: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard

Attendees "Name (Affiliation)":

Apologies:Reggan

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Write ADR about reserved namespace policy	Nathan	Done

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Review of property names: see #PR#12 ● names ● sequences ● lengths Notation for reserved-semantic properties vs implementation-defined properties Reserved namespace policy for future extension of SeqCol https://github.com/ga4gh/seqcol-spec/issues/47	Tim
3.	Alphabet as inherent property of a sequence collection https://github.com/ga4gh/seqcol-spec/issues/46	Nathan
4.	Finalize the actual “minimal schema” that will be part of the spec.	Nathan
5.	Finalize decision on “names-lengths” attribute. (https://github.com/ga4gh/seqcol-spec/issues/40) - name of attribute - required status	Nathan
6.	Finalize ADR: Internal identifiers SHOULD NOT be prefixed https://github.com/ga4gh/seqcol-spec/pull/42	Nathan
	A.O.B.	

Meeting recording:

Meeting minutes:

- Reserved namespace policy
    - need name of properties; during Connect there was a question whether or not we have different nomenclature for properties with reserved semantics and those that do not have reserved semantics.(Issue 47)
    - important discussion because they will be part of digested strings, and part of identifiers. would break backward compatibility.
    - Tim: happy to not change anything but open for discussion

  - Saying anything with underscore would be reserved, would have an entire namespace.
  - May interfere, problem to have endpoint start with underscore?
  - Putting them into schema and reserving it that way is sensible, but problem is how do you add new ones and enforce when you add, you are certain they are not being used by another semantic.
  - We can do uppercase.
  - Nathan: Don't see it as a major problem, don't see us adding a lot of new things. One alternative is saying custom ones need to be prefixed with namespace or colon then they are ours. But we can't enforce that, just say it is a recommendation. Or we could prefix ours but don't want to do that.
  - Low enough problem we can just say this is our list and that's it
  - Ability to enforce is limited by the schema we're working with. Already in things in spec planned which means this becomes a nonissue.
  - Sveinung: agree we shouldn't add underscore to ours, it should be others.
  - Add another custom layer?
  - Can we provide naming convention so we avoid issues/future clashes.
  - Response depends if we know if there's a clash and in communication with the provider, and there is a social aspect here where we can avoid issue.
  - Sveinung: easy to see if there is custom or not.
  - Robert: be like java, attributes and having official ones.
  - Nathan: either way we could, but we should pick on and not spend more time. Seems like what we should do is say in our minimal schema these are the words we are reserving, anything else is fair game to use but we should add that if you use a custom name there is risk of a future version might want to use that name. Risk is low but something to consider.
    - Group agrees with this approach.
    - Sveinung: should there be a way to specify that these are new attributes vs inherited attributes? Nathan: Should be able to do this in json schema.
  - Andy: in implementation phase, we will want to review attributions that have appeared in implementations and if they should be promoted to core attributes set before going to v1.
- Finalize actual minimal schema
  - names. lengths, and alphabets non collated, others would be collated.
  - alphabet referring to lettered alphabet of what could be used to compose sequence with. Agreement whether this is unique character set existing in the sequence or whether or not it will be using upac code therefore you have following values that could be. Molecule type?
  - There are two attributes here. Use case for alphabet existing, don't lose track that this is about proteins. He doesn't like that refget doesn't differentiate between protein and molecule type,
  - 3 attributes here: alphabet, alphabet domain, and molecule type.
  - We're trying to build functionality so people could add these adn it will work, but don't' know if we will standardize molecule type designations. But that's outside scope.
  - There isn't definition, focus on how schema is being defined
  - is alphabet or alphabet domain is element in the schema?
    - at least alphabet is straightforward. Same kind of logic as length
  - Andy: we should let this fall out of implementation, see whether or not things get used. suggest clearinghouse of attributes and define what they could be as

guideline for first implementations. try it and based on that see if they work and what use cases it solves, and if it can be promoted up to main schema.
- what do we propose as minimal? What we have here. anything that comes along is suggested attributes and needs to be defined.
- But some already in there aren't clear enough yet.
- Masks, priorities and typologies. Took all of them out of minimal schema and wrote an extended schema, imports minimal schema and provides working definitions of these. For v1 of spec it will be minimal schema. Eventually some or all will go to main schema and will require implementation to see what's useful.
- Get draft spec out there for people to implement
- names, lengths, sequences would be minimal one
- must a server support names lengths? or is allowed to not implement names lengths.
- Names is required and implementation may not require user to provide names, but need to reference them, might as well say names is required. Names will exist whether you like it or not. Just put them as required.
- For now, compare functions and __ , and spec includes how to implement them should you implement them.
- logical error in compare function. Need to discuss.
- continue to think of what are high priorities things that need to be solved to get draft out there.

2023-05-31: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": Rob Davies,Sveinung Gundersen,Shakun B,Reggan Thomas
Apologies:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Continue draft of spec	Nathan	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	The `/collection` endpoint should provide ability to retrieve either canonical or the concise forms of the seqcol. But how should it be specified? Decision: "Expansion level" specified with "?level=..."	Nathan
3.	Specification draft - https://github.com/ga4gh/seqcol-spec/pull/44	Nathan

	-	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/dXG6PcFgqJ-hSIAdIKS7e3STRS6zM3f0ivlyAhkKSrbarudYFpcPJBIf5WYmhsQM.0V5RdrANL30tTQB-

Meeting minutes:

NS - https://github.com/ga4gh/seqcol-spec/blob/spec_rewrite/docs/specification.md

NS - Discussed about the Terminology streamlining i.e how do we refer to the various levels of seqcol representation?
- Discussed is updating the specification .the goal here is to provide a general, one-stop description of seqcol

NS - path parameter for how to specify the different forms.
NS - the form that you consider the way the sequences are in digest form, but everything else is full. Yeah. That's canonical. And then the concise form was when each attribute of the sequence collection is just a digest.
NS - what we discussed point two on May 17. Was this terminology issue where we were bringing up the names and we made the decision to go back to numbers, right. Okay. The reason why is that the original concern with the numbers was that if we were doing something with pen genomes, then the hedge genome would now be a negative one. Okay, but then we determined that this is actually not an issue, because the pen genome would have to have a different endpoint. Yeah. And so basically, you'd have a slash collection and slash pen genome. So at collection level zero is clear. Yeah. And that pen genome level zero is clear. It's just the top level. So the fact that the numbers just mean the number of lookups relative to whatever it is that I'm whatever endpoint I'm talking about. So like, yeah, for a pen genome level zero, is this level one is this a level one at the pen genome? Is it a set of the things that are level zero at the collection? Yeah. And then maybe that would be like an attribute or array

NS - to continue working on the draft specification

2023-05-23: Reverse Lookup discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Rob Davies,Beatrice A,Haroune Hassine
Apologies: Shakun

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Raise with TASC the documentation of existing prefix and process for creating new one	Tim	

2	Create a tag of the compliance suite for v1.0 before updates	Tim	
3			
4			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Product approval: 4. PRC meeting outcomes: Refget V2.0 - PRC- Comments tracker.xlsx 5. Public comments tracker: Refget V2.0 - Comments tracker.xlsx 6. REWS and DSWS: Refget API_V_2.0 REWS-DSWS-issue-tracker.xlsx	
4.	Compliance report and Compliance document are out of date	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/QG5kHnd5GXLm5sbvRTlxpC0zKtmbFEqjaGyfKEu_rbRwLPt6GDBbaUU-vKRATJc5.2Na47OSfFRbvQKww

Meeting minutes:
TC - Discussed on the outcome from the PRC meeting
TC - Discussed on Compliance report and document

2023-05-17: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Haroune H,Reggan Thomas
Apologies:Shakun, Sveinung Gundersen (Norwegian Constitution Day, national holiday)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Write up ADR and resolve terminology issue	Nathan	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	

2	Terminology https://github.com/ga4gh/seqcol-spec/issues/48 Determined: switch back to numbers, since our original concern is not actually a concern	Nathan
3.	The `/collection` endpoint should provide ability to retrieve either canonical or the concise forms of the seqcol. But how should it be specified? `?form=canonical` (default) and `?form=concise` ? Query parameters means optional. Default is level 2. ?level= is one possibility, ?lookup= but this implies implementation detail, which may not be true.	Nathan
3.	Specification draft - https://github.com/ga4gh/seqcol-spec/pull/44	Nathan
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/f1Lo7yVNCs9FJZ4j6CJh48OBw7Yz7VXise0O7ic_NXKVGQeafPYs9uahy4n1ZHNw.FnrE9Ar6YLrQwtrV

Meeting minutes:

SG: But given that you have a digest of a pan genome, and then you start unpacking, that glitter that he wants to have complete unpacking.
That would mean that you also would start unpacking the actual sequence. collection digest,

SG - digest could be unpacked in different levels also.
TC – Pan genome would be different, either endpoint or different specification, or and at least I'm still on the idea that we would refer to a pendulum the same way we would refer to sequence collection, it just would be referred to at a different level.

TC - sequence collection would be referred to at level one and beyond genome at level zero or sequence production level.
TC - number of lookups and so basically, if you're hitting a pan genome endpoint and you're giving it a zero.
TC - you're given one, then you're going to get back a set of sequence collections. If you hit a collection endpoint and give it a one, then you're going to get back a set of arrays.

NS - level zero is like the digest level one is these arrays and identifiers. And then level two is sort of what we think of as the canonical representation. And then level three would be like the exploded representation that actually has the sequences from the refget server,

NS - a pan genome level zero would be the digestive, the pan genome a collection at level zero would be the digestible collection. pens, you know, a level one would be this compact representation of a pan genome, which is an array of collections. a collection at level one packed representation of a sequence collection. which is a set of arrays of sequences.
AY - if you have a universal endpoint that can take anything.
AY - also have an endpoint that can give you back in different forms, but you better not be able to do both. You have the same endpoint. it's kind of like either you fix it so that it's giving you back the same kind of thing and then make it universal
AY - sort of liked the idea of fixing the endpoint

TC - question of whether we would prefer numbers or names.
TC - humans that would interact with the, with the endpoint.
AY - What is the correct behaviour for the endpoints to do?
AY - I think it would just be an explosion of that JSON payload.
NS - are we going to come up with a different name for every possible I mean, the overload the word explode, the number of possible forms is exploding as you start to add in new other elements that have varying levels of recursion in terms of their identifiers

NS - how do we want to specify the ability to retrieve the different forms to the endpoint,
For example, if you hit the slash collection slash identifier, what do you get back?
You get back the contact representation, or do you get back to canonical representation? And how do you modulate?
NS - is it a query parameter or a path parameter?
AY - if it's a query parameter, you would say question mark. Then what do we use form? Structure?
NS - going back to numbers and I mean, it'd be just slash ID slash number. Gets so could be a path parameter.
TC - I don't even think the query parameters should be mandatory. And, and we should just specify as recommended or as an optional
TC - It makes more sense to have it as a query parameter.
NS - Pan genome defaults to level one. Just because to me, that's the canonical way I think of pans, you know, but you could say, level equals two to get the level one sequence collections, or even level equals three on the pants. You know, those seem both useful.
NS - not everybody who's using the seq collection things is doing because they want to compare genomes, they might just be using it for reproducibility or something like this, where all you really need is given refget server. There's no comparison there among sequences. it's just is, is that the sequence used, then here's its identifier and so the same way I I can't see a lot of use cases
AY - Nathan on what you just said there as well if we want to consider that there are levels of implementation around service collections. So when you said that, the majority use case may not be a case of someone wanting to do the comparison function. is that something we want to say that array a sequence collection implementation may choose not to implement comparison.
NS - query parameter - optional?
AY - still need something in front of it to actually do error checking of basically payload error checking
AY - I would suggest that there's probably still a requirement for a front end server so I will still be okay with query parameters and that front end server

2023-05-03: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Rob Davies,Sveinung Gundersen ,Reggan Thomas,Tim Cezard
Apologies:Shakun

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
	Summarize new terminology proposal in github issue https://github.com/ga4gh/seqcol-spec/issues/48	Nathan	Done
	Review 2 ADRs	Everyone	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2	GA4GH Connect Report/Discussion Key takeaways and next action items Meeting recording	Tim/Andy/Nathan
2.	Previous Actions Review - https://github.com/ga4gh/seqcol-spec/pull/45	Nathan
3.	Specification draft - https://github.com/ga4gh/seqcol-spec/pull/44	Nathan
4.		Nathan
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/lwvk9F--MvbTwNHR6mv3SvNvyKAOW-skW8oqzw1lkPSSeMvqbllX9IbiPGwASMsX.JHO9061CZn5Qhm13

Meeting minutes:
TC -
https://github.com/ga4gh/seqcol-spec/blob/master/docs/decision_record.md#2022-10-05---terminology-decisions

NS - discussed the key takeaways from the connect session.
NS - 1) One of the key takeaway - what the format of the collection that we returned from the server
Currently we are returning the structure that we use in the digest
Return the structure in a different format or make it as an option.
2) ADR sequence types for alphabets

3) How should we reserve attribute names- came up may be because of other ga4gh use cases.

Action items - why we are doing the array based thing. tried to answer this.

We have write up on this, but need to put up those things together.
RD - Big files, which essentially does the sam header, line per reference.
You dont have to pivot the data from array based format. Certain data wont fit in the model very well.

2023-04-25: Reverse Lookup discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": ,Rob Davies ,Reggan Thomas , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Allow namespaced or non namespaced identifiers PR#4	Merged
4.	After PR#39 ENA fails test_metadata_insdc because insdc is not always returned in the metadata payload See query with md5 vs query with insdc	Remove the test for now. Add TODO to clarify why it is removed
5	Continue to look for a 3rd member of PRC ● W Refget API v2.0-REWS_DSWS_questionnaire.docx ●	
6	Compliance matrix	
	A.O.B.	

Meeting recording:https://us02web.zoom.us/rec/share/00WrXSf2dweABzn779VAUwfJJn7cL_cYh1PAa07F77sN7L9dFd2pb9sQzVI8jE-X.GZfobHaRizpWmij7

Meeting minutes:

2023-04-19: Cancelled (due to Connect meeting)

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

2023-04-05: Connect prep and ADR

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Shakuntala Baichoo ,Sveinung Gundersen ,Reggan Thomas , , ,
,
Apologies - Rob Davies

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Preparation for the GA4GH connect presentation • Slides	
4.	Specification of the service info	
5.	Do we need an ADR for the specification of the collection endpoint ?	
	A.O.B.	

Meeting recording:https://us02web.zoom.us/rec/share/-CADNH_Mw1ZoaTnTKkyFVDqBXqIT_Z0GzDzDC4C_ul822cDjoGyiO5cAegSSz4_q.PgjEILEAfT_BgTI5

Meeting minutes :

2023-03-28: Reverse Lookup discussion – CANCELLED

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
	A.O.B.	

Meeting recording:

Meeting minutes;

2023-03-22: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Rob Davies ,Sveinung Gundersen ,Shakuntala B,Chen Chen , ,Reggan Thomas (GA4GH)
Apologies: Andy

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Write up ADRs on decisions: - Seqcol will RECOMMEND implementing names-lengths - Seqcol will REQUIRE implementing inherent attributes	Nathan	
2	Put together some slides	Everyone	2 weeks

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Question 1: Terminology for “inherent” Outcome: we like ‘inherent’	Nathan
4.	Question 2: How should the compare function handle names_lengths_sorted?	Nathan
	A.O.B.	

Meeting recording:https://us02web.zoom.us/rec/share/eZufvbVtQQ1vi4VsGVsPbsPjlVAEjdftT5aiVBpaKokcRBKxnrP3Y53hUozfIWMc.sc0bAKJesT6TrzXL

Meeting minutes:
NS - https://seqcolapi.databio.org/comparison/59319772d1bcf2e0dd4b8a296f2d9682/2e7bc302a54ecec62d8155e19fbf2748

2023-03-08: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": ,Rob Davies ,Sveinung Gundersen , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Change connection session to april 19 from 3-4pm	Reagan	not needed after all
2	Write up ADRs on decisions: - Seqcol will RECOMMEND implementing names-lengths - Seqcol will REQUIRE implementing inherent attributes	Nathan	carry

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review Your GA4GH April Connect Session proposal, Sequence collections: How to uniquely identify and compare genomes, has been approved, and has been scheduled for Friday April 21 at 13:30 - 15:00 BST . Please let us know if this time/date no longer works for you by **Tuesday March 7**, otherwise we will proceed as scheduled. Registration is open now: https://broadinstitute.swoogo.com/ga4ghaprilconnect23/	
3.	Continue discussion on undigested, non-1-to-1-attributes, and names-lengths array https://github.com/ga4gh/seqcol-spec/issues/40 Look at examples of names-lengths: Function: https://github.com/refgenie/seqcol/blob/46675b669ae07db9da4fc3d113fefa2c1667b1fb/seqcol/seqcol.py#L300-L310 handling inherent attributes: https://github.com/databio/henge/blob/b53333072a51a0bd532f14d3ed949dd02e67d6be/henge/henge.py#L376-L383	Nathan
4.		
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/3-7hZTWkpbc63MMYbgfSmf5bkv3HD4p6NqAAjJaf7TnuxDW439dd_mI7kP9Xa5ze.mmrXHNZE8OsnLaBK

Meeting minutes:
NS - https://seqcolapi.databio.org/
NS - https://seqcolapi.databio.org/comparison/59319772d1bcf2e0dd4b8a296f2d9682/2e7bc302a54ecec62d8155e19fbf2748

2023-02-28: Reverse Lookup discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Rob Davies,Reggan Thomas (GA4GH)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Continue to look for a 3rd member of PRC	Tim	
2	Look at reporting on the compliance suite	Tim	
	Discuss with ENA about implementation of 404 instead of 400	Tim	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review ● Change the text around prefix to enforce support of resolution with or without namespace PR#4 Additional work required to update Perl code and ipython notebook ● Fill in Forms ○ **Need 3rd PRC member** ○ W Refget API-v2.0-product_approval_submission_form.docx Completed ○ W Refget API v2.0-REWS_DSWS_questionnaire.docx Some feedback from secretariat ○ W PE _ Comms Intake Form - Refget V2.0.docx Completed	
3.	Questions: ● Should support for MD5 be mandatory? This line says it is optional but this line says it is. <- Clarified	

	- This line advertise the out of date Refget compliance report - Different interpretation of <400 Bad Request> and <404 Not Found> https://www.ebi.ac.uk/ena/cram/ assumes it receives a md5 and if it has an incorrect length it reports <400 Bad Request> https://refgetv2.fly.dev/ reports an <404 Not Found>	
4.	Compliance tests updates: PR#39 - Relaxed compliance for 400 errors - Disabled Squid sniffing - Add Github action support	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/share/etbBqzqq-RDV_36jn5nL3w18Ior-hYIcRzQo7iMx8PPR12Zc7N1o-rKQHGS5ym_U.v6kUPWMlkRbtHcoJ?startTime=1677603839000

Meeting minutes:

2023-02-22:Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Sveinung Gundersen ,Chen Chen ,Rob Davies ,Reggan Thomas , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Start to put together some paper text	Nathan	two weeks
2	Submit Connect proposal	Nathan	Done Feb 22

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Finalize decision on connect session Will require: - presentation of use cases, - justification for choices we made, - ideally a draft of specification (but not absolutely required) - some advertising or outreach beforehand to people who may be	

	interested Sequence collections: How to uniquely identify and compare genomes April 20, afternoon	
4.	Continue discussion on undigested, non-1-to-1-attributes, and names-lengths array https://github.com/ga4gh/seqcol-spec/issues/40	
	A.O.B.	

Meeting recording:
https://us02web.zoom.us/rec/play/iuDleYXf0j4YgWnn5uuoO1udVzNE-nBr23RL8oS6fRjX6cjNIw_3bXGir2wWCr3Xt0w6bkysGRtoE9WB.C4f1IzqqgSxxHtN5?continueMode=true&_x_zm_rtaid=ZIl9Q4bMSkyXzrLTSSH1tA.1677077849319.4fd94e45beeaedd5693060223191451d&_x_zm_rhtaid=674

Meeting minutes:

2023-02-08: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Robert Davies ,Sveinung Gundersen ,Reggan Thomas , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Start to put together some paper text	Nathan	Carry
2	Finalize decision on connect session		Next time.

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	GA4GH connect session proposals: https://docs.google.com/forms/d/e/1FAIpQLSf0c5FMnzeZFEU3-0aU4goh5qFL404U8KQZZMiXF-W0tikeIg/viewform Two possible proposals: 1. introducing/showcasing sequence collections (are we ready?) a. File formats (BED, VCF), VRS, Beacon?, Phenopackets?	

	2. prefix discussion? probably not	
4.	Discussion on undigested, non-1-to-1-attributes, and names-lengths array https://github.com/ga4gh/seqcol-spec/issues/40	
	Issue of building the identifier with a prefix should be discussed with GKS and VRS.	
	A.O.B. Review GA4GH Product Graphics (Feedback by Friday, February 10) ● the graphic and the graphical representation of the concepts ● the text in the graphic ● Caption	Deadline - Friday, February 10

Meeting recording:https://us02web.zoom.us/rec/share/GV_F0njS20-GRArjCSDQIhnV53-uJecvDkfVW7HNWQzxWwuLpc73GtwI3fI6Agq-.1IgW1OQG4umpJSNH?startTime=1675868378000

Meeting minutes:
NS - https://github.com/ga4gh/seqcol-spec/pull/32/files

2023-01-31: Reverse Lookup discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: Shakuntala Baichoo (Uni of Mauritius)

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Change the text around prefix to enforce support of resolution with or without namespace	Tim	
2	Fill in Forms	Tim	
3			
4			
5			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	

3.	GA4GH:SQ. prefix for ga4gh identifier ? - Sha512 same size as an md5 so complicated to recognise Options url: - http://refget/sequence/ga4gh:SQ.NNNNNN - http://refget/sequence/SQ.NNNNNN - http://refget/sequence/refseq:NM_00000.1 - http://refget/sequence/MD5NNN - http://refget/sequence/md5:MD5NNN	Tim
4.	What should we do about the compliance suite website ?	Tim/Andy

5	Refget approval:
	- Refget API-v2.0-product_approval_submission_form.docx
	  ○ Choose PRC committee members?
	- Refget API v2.0-REWS_DSWS_questionnaire.docx
	  ○ Use case presentation
	  ○ Milestones of tasks to complete before approval
	  ○ Layman summary of Refget
	  ○ Review list of risks associated with Refget
	- PE _ Comms Intake Form - Refget V2.0.docx
	  ○ Do we need to update the cave entrance ?
		Tim

	A.O.B.	

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	Remove Trun512,Remove naming authorities - PR#3: Reviewed	Tim	
2	Create Security questionnaire for Refget 2.0 from the v1.0: DONE	Tim/Regan	
3	Create issues in https://github.com/ga4gh/approval-tracker/issues https://github.com/ga4gh/approval-tracker/issues/24	Regan	

Meeting recording:
https://us02web.zoom.us/rec/share/5-UvxDxXtxKQ8hONTPo0DjXRJn_miRav0avjrJylfwZ1MIVfBaGvSAWbGAJavPG-.yf-pC4bg2y48zYfX?startTime=1675184399000

Meeting minutes:

Meeting transcript - https://otter.ai/u/vjlBLaNSUpj70sgenNwdJmOMbJE?tab=summary&utm_source=va_chat_link_2_exp

TC: https://www.ebi.ac.uk/ena/cram/swagger-ui/index.html?configUrl=/ena/cram/v3/api-docs/swagger-config

AY: http://refget/sequence/ga4gh:SQ.NNNNNN
AY: http://refget/sequence/SQ.NNNNNN
AY:http://refget/sequence/refseq:NM_00000.1
AY: http://refget/sequence/MD5NNN
AY:http://refget/sequence/md5:MD5NNN
AY:https://www.sanger.ac.uk/person/aleksandra-ola-tarkowska/

2023-01-25: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: , Rob Davies, Chen Chen, Sveinung Gundersen ,Reggan Thomas ,Andy Yates , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	New issue to be raise from Sveinugu’s comment	Sveinung	Nevermind
2	Issue of building the identifier with a prefix should be discussed with GKS and VRS. Andy to reach out to Alex re: VRS CURIES	Andy	
	GA4GH connect session proposals: https://docs.google.com/forms/d/e/1FAIpQLSf0c5FMnzeZFEU3-0aU4goh5qFL404U8KQZZMiXF-W0tikelg/viewform	To do next time	March 1st
	Comment on Tim’s GSOC draft (GSOC proposal)	Everyone	Done, submitted
	Comment on new “names-lengths” array proposal: https://github.com/ga4gh/seqcol-spec/issues/40	Everyone	Done

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	Nathan
2.	Previous Actions Review	Nathan
3.	Go over almost-done ADRs: - Finalize PR #34 ADR for serialisation of sequence collection - OK - Get approval for #41 ascii array name recommendation - OK	Nathan/Tim

	- https://github.com/ga4gh/seqcol-spec/pull/42 - WAIT for discussion	
4.	/service-info vs /metadata - https://github.com/ga4gh/seqcol-spec/issues/3 - https://github.com/ga4gh/seqcol-spec/issues/39	Nathan/Tim
5.	Discussion on undigested, non-1-to-1-attributes, and names-lengths array https://github.com/ga4gh/seqcol-spec/issues/40	Nathan/Sveinung
	A.O.B. Sequence collection implementation GSOC proposal	Tim

Meeting recording :
https://us02web.zoom.us/rec/share/mqsNhXcqTWJzbTaZRslO7d6N-3_IcIJRzXO3JMp0p8ruhFFV42-s4SJGcBPw28Sy.IIC-dV3NsKwaITuA

Meeting transcript:
https://otter.ai/u/NkJFX7GPnUOiO4Y3WgFVpFy4T4s?r=&i=0&f=oIn&u=true&utm_source=va_chat_link_1_exp

Meeting minutes:

2023-01-11: Refget discussion

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”: ,Shakuntala Baichoo ,Rob Davies ,Chen Chen ,Reggan Thomas (GA4GH) ,Sveinung Gundersen ,Rasko L , Andy Yates

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	• Issue of building the identifier with a prefix should be discussed with GKS and VRS • Is there plans for GA4GH to resolve CURIES that exists in refget and could exist in seqCol? Answer: No. ○ Andy to reach out to Alex re: VRS CURIES ○ Refget to allow non-prefixed identifier lookup – In progress	Andy	
2	New issue to be raise from Sveinugu’s comment	Sveinung	carry
3	Review PR#34, to be merged in 2 weeks.	Everyone	By next time

4	Write a new issue about non-digested arrays idea	Nathan	Done #40
5	Issue #33: Create an ADR to formalise the set of characters accepted/recommended in array names	Nathan	Done #41

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Issue #8: What information is included within the string-to-digest? ● Which field should be mandatory ? ● What are the reserved field names	
4.	Issue #39: Service info specification	
5.	Issue #37: ● 37A: Should we prefix things internally? to be closed by PR#34 ● 37B: Should we prefix the final level 0 digest in what we refer to as the "seqcol identifier?" To be discussed. Maybe depends on: ○ Issue of building the identifier with a prefix should be discussed with GKS and VRS ○ Is there plans for GA4GH to resolve CURIES that exists in refget and could exist in seqCol ? ○ We should kick this can down the road, it is not critical for now, we can just not prefix and then see what the conclusion is	
	A.O.B.	

Meeting recording :
https://us02web.zoom.us/rec/share/h5EzZ6inbcP6gtj7hApgryHymZ-ejUQpJEp2dfJEJLb0qavu9EOKpGY4DSO4c_8m.TG-5cJxmzhKy2kbv

Meeting minutes:

NS - issue 33 - you must at least use UTF eight but but we recommend that you just restrict it to ASCII
if you don't want to you're just going to lose interoperability than with other services that are going to expect it to be ASCII
NS - PR 31 - Andy has addressed the comments - In progress
Action item 3 from the previous action items - Andy to look into it.

TC - https://github.com/ga4gh/ga4gh-registry/pull/7

TC - Create issue for Sequence collection’s service info specification linked this issue to is the discussion about the schema that we should, that we should be used to specify what's going to be inside the sequence collection that will be provided by this particular implementation.

TC - PR 34 -Done. except a comment - how the the sequence production is going to be serialized and it states that we are going to use the RFC
we do need the sorting of array names if people don't use ASCII.
the second point is why do we not need floating point? Do we have any floating point data?

TC - The only advantage of using RFC 8785 was to have a much simpler specification. So you don't need to specify what you use or what you don't use. you just say we have the standard that is going to provide the serialisation method
SG - I think one thing, because if for instance, you don't want to have a external library, dependency and so on, then it's very easy to implement the parts that we need
TC - worth putting that Python function in there. Just like here's how you could implement the basic case that satisfies the default settings in Python without an external library in a three line function

NS - review PR 34
TC- Service info -the main piece of information that needs to be part of service info, is what to expect in that particular search sequence collection implementation in terms of the supported array names.

NS - maybe in version 1.1, or whatever, we we could add specifications for these optional features that are not part of the original specifications. And I think when we add those, then it would make sense for us to also add a classification in the service info that we implement that optional standard.
NS - we should focus on is basically finalizing that default schema. - in next 2 weeks

2023-01-03: Reverse Lookup discussion - CANCELLED

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees “Name (Affiliation)”:

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			
3			
4			
5			

	Agenda Item	Person/Time

1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1	● Issue #33: Create an ADR to formalise the set of characters accepted/recommended in array names ○ Option 2A, MUST use UTF8, SHOULD use ascii	Nathan	2 weeks
2	● PR#31: Rephrased to focus on Sequence identifier	Andy	Pending
3			
4	● Create issue for Sequence collection's service info specification	Tim	See issue #39
5	● Issue #37: to be closed by PR#34 ● New issue to be raise from Sveinugu's comment		

Meeting recording:

Meeting minutes:

2023-MM-DD: Template

Chair: Andy Yates / Nathan Sheffield / Tim Cezard
Attendees "Name (Affiliation)": , , , , , , ,

	Actions Arising (label SC/RL/general)	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.		
4.		
	A.O.B.	