Global Alliance
for Genomics & Health

GA4GH Future of VCF

Minutes and Actions 2023 - 2026

Previous minutes: 2018 | 2019 | 2020 | 2021 | 2022

These are the minutes for the File Formats meetings, a subgroup of the GA4GH Genomic Knowledge Standards (GKS) Work Stream. For further information, please visit the GA4GH Work Stream page.

Important documents:
Landscaping and Benchmarking document -
working_doc-Future _of_VCF_Landscape_Analysis-211006

Table of Contents

Meeting Protocols
Agendas, Minutes and Actions
- 2026-09-01
- 2026-05-12
- 2026-03-17 : Blogpost
- 2026-01-20 : Previous action items
- 2025-11-25 : Review landscape analysis
- 2025-09-30 : VCF Zarr
- 2025-06-10 : Landscape analysis
- 2025-03-18 : Gnome AD
- 2025-02-18 : Roadmap 2025
- 2025-01-21 : Cancelled due to unavailability of the leads
- 2024-11-26 : Connect planning
- 2024-10-29 : Public data set
- 2024-10-01: Cancelled (due to unavailability of the leads)
- 2024-09-03: NIH roadmap and VCFcompression challenge
- 2024-05-14: Challenge planning & Publication
- 2024-03-19: Challenge planning & Survey review
- 2024-02-20: PR 434/435 & Benchmark Challenge
- 2024-01-30: Benchmarking and PR434
- 2023-11-28: PR 434
- 2023-10-31: PR 434, Blog post
- 2023-10-03: Plenary and UK Biobank
- 2023-09-05: UK Biobank / Illumina Use Case
- 2023-07-10:Local Alleles & Blocksize

2023-06-12: Connect action items & VCF specification
2023-03-20: Synthetic Geno/Phenotypes
2023-02-20: SVCR data model & datasets
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 All deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd

Agendas, Minutes and Actions

2026-11-17

Chair: Ole Schulz-Trieglaff (Ilumina)
Attendees: Daniel Puthawala (Nationwide Children's), Rob Davies (Sanger), Melissa Cline, Reggan Thomas, James Bonfield, Alex Wagner
Apologies: Albert Smith (Michigan), Beatrice Amos

	Actions Arising	Assigned To	Deadline
1.			
2.			
3.			
4.			
5.			

Link to meeting recording:

Meeting summary:

Global Alliance
for Genomics & Health

2026-10-27

Chair: Ole Schulz-Trieglaff (Ilumina)
Attendees: Daniel Puthawala (Nationwide Children's), Rob Davies (Sanger), Melissa Cline, Reggan Thomas, James Bonfield,, Alex Wagner
Apologies: Albert Smith (Michigan), Beatrice Amos

Actions arising:

- ☐ **Ole** - Re-ping Genomics England (head of bioinformatics engineering) about presenting.
- ☐ **Melissa** - Arrange a pangenome / graph genome speaker; confirm presenter.
- ☐ **Ole** - Continue editing the blog draft; share it with Melissa for a fresh-eyes edit when ready.
- ☐ **Group** - Decide blog post vs commentary once Daniel Cameron has commented on the draft.
- ☑ ~~**Beatrice** - Circulate comms team feedback on the blog post request.~~
- ☑ ~~**Melissa / Beatrice** - Add Ole to the mailing list; review list, Slack and meeting invite membership.~~
- ☑ ~~**Chair** - Add think tank projects to next agenda; consider ending calls 10 min before the hour.~~

	Agenda Item	Person/ Time
1.	Welcome, Introductions, New Faces, & Confirm Agenda ● Reminders: ● Think Tank projects: https://github.com/ga4gh/gks-portal	
2.	Review of action items from the previous meeting	
3.	Use cases document (feedback round)	
	Blog post vs peer-reviewed commentary (listed by Beatrice as "reviewing the article repository")	
	Think tank projects (deferred)	

Link to meeting recording:
https://us02web.zoom.us/rec/share/HP13GQLU35_5GWKce8ybML9G5cTkZXPx2FYRfysEYVdwYFUfpRcnHycSKJc_klBz.jAV0_xy2IxDXlAFd

Global Alliance
for Genomics & Health

Meeting summary:
Future of VCF paper

The group will keep the blog post on its original track (a "call to arms" for the group) unless it can be reframed around impact, and will bring in Daniel Cameron, the VCF spec maintainer, before finalising any write-up.

- **Guest speakers:** Melissa will line up a pangenome / graph genome speaker; Ole will re-ping Genomics England; James may present his earlier work once he picks it back up.
- **Blog vs commentary:** a peer-reviewed commentary would need a view on impact (which ideas will move forward); the current draft lists problems without solutions. Melissa offered editing support.
- **Publication precedent:** the GKM commentary has cleared external review at Nature Genetics, showing a route for GA4GH work.
- **Daniel Cameron:** James will invite him; his input would add weight to any article. Next Australia-friendly call is Tuesday 17 November.
- **Admin:** shared drafting folder exists; Ole needs adding to the mailing list; Melissa and Beatrice to review list/Slack membership.
- **Format:** proposal to end calls 10 minutes before the hour, which Ole welcomed.

Review of previous action items

Most previous items are done or in progress; two need follow-up (blog co-writers, list membership).

#	Item	Owner	Status	Notes
1	Coordinate with Daniel Cameron / GA4GH members in Australia	Ole, Albert, Beatrice	Done	Late-UK-evening calls were cancelled (9pm hard for Ole and Albert). Daniel wants to stay involved; agreed quarterly Australia-friendly calls. Beatrice sent invites; next is Tue 17 Nov.
2	Blog post comms request	Beatrice	Submitted	Comms response expected 1 Nov. Ole unclear what the outcome would be; Reggan understood comms will review and return corrections, which Beatrice will circulate.
3	Guest speaker pipeline	Ole, Rob, others	In progress	See Guest speakers section.

4	Close out use-case document feedback	Reggan	In progress	Feedback received on both shared use cases; link to be added to the agenda doc.
5	Shared access for blog editing; recruit co-writers	Ole	Half done	Draft is shared; no co-writers yet — Ole is the only contributor.
6	Share correct Google Doc link for drafting workspace	Daniel Puthawala	Done	Folder exists; Melissa posted the link in chat.
7	New Slack channel for expanded Work Stream Leads group	Melissa / Beatrice	Done	GKS Product Leadership channel exists (created by Beatrice); Albert, Ole and James are members.
8–10	Invite guest speakers	Various	In progress	Folded into speaker discussion.
11	Share attendee emails so Melissa can add them to mailing list	—	Done	Ole reports he is still not on the list (did not receive Reggan's use-case email).

Use cases under review:

- **Merging samples from multiple VCFs** - feedback from Oliver on a tentative merge workflow and combining with other use cases.
- **Normalised representation of variants** - several comments received.

Guest speaker pipeline

One speaker is now likely (pangenome / graph genomes via Melissa); the others are pending.

Candidate	Lead	Status
Pangenome / graph genome team (Human Pangenome Reference, a GA4GH Driver Project)	Melissa Cline	Melissa spoke with Karen Miga; the team is willing and just needs to choose a presenter. Much of the work is at Melissa's institution. Ole welcomed it.
Genomics England - head of bioinformatics engineering	Ole	Emailed with background; no reply yet. Ole will ping again. Interest: large VCF-based cohorts and infrastructure.

James Bonfield - his earlier VCF-related work	Rob suggested	Work paused for about a month after discussions with Rob. James may present "when I get back into it"; no date. Rob clarified he supports the idea but specifics need more discussion.

- Melissa noted the graph genome topic differs from the scalability focus the group has had so far.

Blog post vs peer-reviewed commentary

No final decision: the leaning is to finish the blog post as planned unless the piece can say which ideas are likely to move forward, which a commentary would need.

Ole's framing of the current draft

- Proposed early this year to update an earlier blog post, publicise the group and attract more people to the calls.
- Built from Reggan's use-case spreadsheet: who the group is, current VCF problems, what any evolution or replacement should deliver, and past guest talks (Broad Institute; a VCF-Zarr talk.
- Ole described it as "a list of complaints" without solutions. He is fine with a commentary but unsure how to change tone and scope.
- Ole's time is limited; he hopes to edit more towards year end. He pasted the draft text into the chat.

Melissa's input

- A commentary would need a sense of impact - which ideas are likely to progress. If the draft can't speak to that, polish and publish the blog post ("a call to arms").
- Albert said on the previous call that the draft's content has already motivated recent hackathon topics, so the ideas have community value.
- Offered to edit and give a fresh read once Ole feels it is ready.

GKM precedent (Nature Genetics)

- Melissa reported the GKM commentary (covering VRS, VA-Spec and Cat-VRS, grouped as the Genomic Knowledge Model) has finished external review; some editor comments remain.
- Submitted at 1,200 words; reviewers asked for ~1,500 words of additions, and the editor invited a 2,000-word resubmission.
- Any new commentary would still need a 1,200-word initial submission. The editor seemed enthusiastic about GA4GH work; other venues exist.

Drafting workspace

- Daniel Puthawala's shared folder holds a placeholder "Future of VCF Article" Google Doc with tabs, including the GKM Nature Genetics article and its cover letter as optional structural templates, plus a blank tab to start fresh.

Engaging Daniel Cameron (VCF spec maintainer)

James will invite Daniel Cameron, since the group risks drifting from the actual VCF specification without its sole maintainer.

- James raised an open scoping question: does "future of VCF" mean fixing the current spec as an interim measure, or what replaces VCF? Both were on the table in the past.
- Interim improvements such as local alleles have already gone in; it's unclear whether other simple fixes remain.
- Daniel's contribution would add weight to any blog post or commentary.
- He is in Australia, so an evening (UK) slot is needed; James doesn't want to require very late attendance. The quarterly Australia-friendly call (next: Tue 17 Nov) is the natural fit.
- Melissa (Pacific time) said Daniel is the more critical attendee for that slot; she can follow via minutes.

Mailing list, Slack and GKS leadership meetings

Membership is patchy: Ole isn't on the mailing list, and James finds the post-merger channels and invites confusing.

- **Mailing list:** Ole did not receive Reggan's use-case email and asked to be added.
- **GKS Product Leadership Slack channel:** created by Beatrice; Albert, Ole and James are members. James rarely opens GA4GH Slack.
- **GKS product leads meeting series:** James gets invites (e.g. "GKS Leaders") but has never been sure he's meant to attend. Melissa explained the old series covered the three pre-merger GKS standards; a single cross-time-zone leads call proved impractical and tended to become round-robin status updates. Instead, GKS leads (Melissa, Alex) now join work-stream calls such as this one. James does not need to attend the leads meeting.
- Melissa will review with Beatrice who is on which list to cut irrelevant invites.
- Rob asked that the shared links end up in the minutes.

Deferred items and meeting format

- **Think tank projects:** deferred to the next call - Melissa felt it couldn't be covered quickly and people were leaving for other meetings.
- **Meeting length:** Melissa suggested ending 10 minutes before the hour, as GKM calls do, to avoid overruns and give transition time. Ole agreed it's a good idea.
- **Next meeting:** next month (regular slot); Australia-friendly call on Tue 17 Nov.

Global Alliance
for Genomics & Health

2026-09-01

Chair: Ole Schulz-Trieglaff (Ilumina), Albert Smith (Michigan)
Attendees: Daniel Puthawala (Nationwide Children's), Jonn Smith (Broad), Rob Davies (Sanger), Wallace Minxian Wang (Beijing), Melissa Cline
Apologies: Beatrice Amos

	Actions Arising	Assigned To	Deadline
1.	Reach out to Daniel Cameron: how to coordinate with GA4GH members in Australia?	Ole, Albert, Beatrice	Next meeting - Completed
2.	Blog post draft: how to share this? Ole to coordinate with Beatrice	Ole/Beatrice	Next meeting Submitted comms request. Expecting to hear back on 1 Nov Bea to report back
3.	Guest speaker pipeline	Jonn, Rob, Ole to follow up	Next meeting - In progress - Ole to ping GE. Also, check with Pangenome group (Karen Miga)
4.	Check with Reggan on use-cases document feedback; close out if none	Ole	Add link here
5.	Set up shared access for blog post editing; recruit co-writers	Ole	Blogpost - Done. Recruit - in progress
6.	Lead next steps on a potential commentary-style article; share Nature Genetics submission as a template	Melissa	
7.	Fix/share correct Google Docs link for the drafting workspace; post to Slack and email	Daniel (Puthawala)	Completed.

8.	Ask Beatrice to create a new Slack channel for the expanded Work Stream leads group	Melissa	Completed
9.	Consider reaching out to the GVS/All of Us team about a guest talk	Jonn	In progress
10	Consider inviting James (Sanger) to speak on zstd/BGZF replacement work	Rob	In progress
11.	Reconnect with Genomics England bioinformatics contacts re: guest talk	Ole	In progress
12.	Share email addresses in chat/with Melissa to join the mailing list	New attendees	Completed

	Agenda Item	Person/ Time
1.	Welcome, Introductions, New Faces, & Confirm Agenda • Reminders: **GA4GH Plenary registration** open: https://broadinstitute.swoogo.com/ga4gh-14th-plenary/11058020 • Think Tank projects: https://github.com/ga4gh/gks-portal	
2.	• Meeting logistics: new time ok for everyone? How about the AUS friendly meeting? • community feedback for the two priority use cases. Reggan initiated this in December 2025, I don't remember if we heard anything back? • Current status of blog post : spoiler, I had almost no time to work on this since the last meeting. Current draft is attached. • Any other items? • guest speaker pipeline: anything we can think of?	
3.	A.O.B	

Link to meeting recording:
https://us02web.zoom.us/rec/share/z6hRbbIrCMuvZAOPTIIC1I4u7XbMIXmhVnHhw43uEuWIAYptZVfOguDwzO4f-aoG.WuCGr2fxxDIX4C0i

Meeting summary:
Future of VCF Commentary / Review article repository

Meeting Logistics

The new meeting time (one hour later than before) seemed acceptable given higher attendance. Rob noted the current VCF spec maintainer, Daniel Cameron, is poorly served by any of the current times.

The group previously tried an "Australia-friendly" alternate meeting but had to cancel it since neither Ole nor Albert could attend.

🔧**Action item:** Ole will reach out directly to Daniel Cameron spec maintainer, cc'ing Beatrice, to find a workable arrangement.

Use-Cases Document (Leftover Item)

A prior document collected use cases that any VCF replacement should satisfy for large cohorts. Feedback was solicited from the mailing list (via Reggan) around last December, but no responses came in, and Reggan has since left the group.

🔧**Action item:** Ole will check with Reggan on any feedback received; if none, the item will be closed as "attempted but not pursued further."

Blog Post Update

Background: a blog post (~2.5 years old) describes the working group's purpose, guest speakers, and achievements (notably contributing to the VCF spec). About 60% of attendees weren't familiar with this history.

Ole has a draft update summarising recent guest speakers and results, but limited time has slowed progress. He shared a link to the current live blog post in the chat and plans to work with Beatrice on a shared folder so others can co-edit.

Goal: Publish the updated post this autumn.

🔧**Action item:** Ole to set up shared editing access; open call for volunteers to help write/edit.

Publication Strategy Discussion (major topic)

A substantial part of the meeting was devoted to whether/how to formalise the group's work in a publication:

Albert noted the group has previously discussed whether there's "enough meat" for a full research publication, and has generally settled on the blog post as more realistic given time constraints.

💡Daniel suggested a commentary-type article - shorter, still peer-reviewed, positions the group as representing the VCF standard, summarises recent developments, and issues a "call to the community." He noted this is a lower bar than a full research article and offered to share tooling/templates his group built for consortium-style paper writing (a Google Docs–based collaborative workflow plus a co-author/contribution tracking spreadsheet aligned with CRediT criteria).

Jonn strongly supported a commentary/review article for now, noting that without a formal publication, the community rarely learns about the group's work even if blog posts exist. He suggested reserving a full research paper for a future major milestone (e.g. a "VCF5" with breaking changes).

Melissa shared that GKS already has a similar commentary article in review at Nature Genetics (a few pages, one figure, ≤10 references) and offered it as a template - she volunteered to coordinate next steps.
Daniel noted his group already has a Nature Biotechnology–style commentary template (from the GKS overview paper) that could be adapted, and set up a shared Google Docs directory with reference copies and fresh tabs for this group to use.
Consensus: the blog post and a possible commentary article aren't mutually exclusive - the group can pursue both, and the blog post draft may translate reasonably well into commentary format.
Communication preference: Discussion (email vs. Slack) was raised for coordinating this work - email was the clear preference of the group (Wallace and Ole both said Slack access is inconvenient for them).

🔧**Action items:**
Melissa to take point on next steps for a potential commentary article draft.
Daniel to fix/share the correct Google Docs link (there was some mix-up with a "VASpec" vs. "Future VCF paper" link) and post it to Slack/email for a permanent record.
Melissa to ask Beatrice to set up a new, broader Slack channel for the expanded set of Work Stream leads once she's back online; Daniel will also post the link there and in meeting minutes.
New attendees not yet on the mailing list were asked to drop their email in chat to be added.

GA4GH Plenary/Connect Meeting (Sept 28–Oct 2)
Melissa asked if anyone planned to attend in person to potentially organise an ad hoc meetup.
Albert wasn't planning to attend (short notice, unclear direct benefit) but was open to it in principle.
The next Connect meeting (6 months out) will be in a location more convenient for European attendees Hinxton, which the group saw as a better target for an in-person meetup.

Guest Speakers
Ole reflected that past guest speakers (e.g. Jerome/Keller on a CRAM-like/VCFSA format, someone from the Hail team) had consistently enriched discussions, even though the group has struggled to generate concrete action items historically.
Proposals for future speakers:
Jonn suggested inviting someone from the GVS (Genomic Variant Store) team working on the All of Us program's cloud-based variant storage format, which addresses scalability limitations of flat VCF files.
Rob mentioned the Sanger/HTSlib "file formats" working group's ongoing discussion about replacing VCF's BGZF compression with Zstandard (zstd)-based compression (and associated index changes) - noted this group's next call is Oct 6. Rob also floated inviting a colleague (James, currently on holiday) to give a talk on this work.

Ole noted possible contacts at Genomics England's bioinformatics team who are thinking about what could replace VCF for large cohorts - a longstanding to-do to reconnect with them.

Idea floated: publish short summaries of guest talks and advertise the call more widely to attract more attendees ("seminar experience").
Historical speaker themes recapped by Albert: VCF size/scaling challenges, various proprietary/solution-specific formats (e.g. the "Savvy" format from his group), difficulties simulating realistic large-scale/rare-variant data due to data-sharing restrictions, and a now-defunct biobank-related project (flagged by Rob as needing formal sign-off/cleanup, which Albert confirmed he already did on his end).

Housekeeping
Daniel set up a shared Google Docs directory with reference files from the prior GKS overview paper and fresh working tabs for this group.
A co-author/contribution tracking spreadsheet (based on CRediT taxonomy, used by Nature/Springer journals) was highlighted as a useful tool for managing multi-author consortium papers.
Ole will add meeting notes into a Word document Rob posted in Slack, since he doesn't have everyone's email address directly.

2026-05-12

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees: Tim Hefferon, Rob Davies, James Bonfield, Lee Lichstenstein

	Actions Arising	Assigned To	Deadline
1.	Continue developing the use-case focused blog post. Aim for a substantially complete draft by mid/late summer to allow GA4GH review before the plenary. Initial draft by mid / late - summer	Ole	Before plenary
2.	BioBank-scale data format paper from Exeter	Albert	Before next meeting
3.	Submit the completed blog post through the GA4GH Comms review process.	Ole / GA4GH support	
4.	Clarify the outstanding action regarding inviting the University of Exeter/biobank-scale data format group, as the original context is no longer clear.	Albert	
5.	Continue work on improving VCF/BCF indexing performance.	Technical lead	

		responsible for indexing	
6.	Evaluate OpenZL on BCF data and report findings back to the group.	Technical lead evaluating OpenZL	
7.	Share the OpenZL library and relevant GitHub discussion with the group.	OpenZL evaluator	
8.	Continue investigating how graph genome representations should be incorporated into future variant normalization recommendations.	Working group	
9.	Assess engineering effort required for integrating the imputation service into meta-imputation workflows, and design a coordinated experiment across reference panels.	Proposal lead, with follow-up involving Albert's group	
10.	Continue monitoring developments in indexing, compression methods and related technologies.	Working group	
11.	Consider inviting additional guest speakers on genomic file formats and emerging technologies.	Working group	
12.	Explore proposing a GA4GH Connect session around the completed blog post at the Singapore Plenary.	Working group	

	Agenda Item	Person/Time
1.	Welcome, Introductions, New Faces, & Confirm Agenda	Ole/ Albert
2.	Previous Actions Review	Ole / Albert
3.	Review blog posts and use cases ga4gh_future_of_vcf_draft_blog_post_2026.docx Landscape Analysis document here	Ole
4.	A.O.B GA4GH 14th Plenary - GA4GH 14th Plenary will take place from 28 September to 2 October 2026 in Singapore at the One Farrer Hotel. Details here: https://broadinstitute.swoogo.com/ga4gh-14th-plenary/11059298	Reggan

	Early bird registration closes on May 14, 2026 Next meeting on June 9, 2026 at 9 pm BST	

eye

	Previous Actions	Deadline
1.	Consider inviting the authors of the BioBank-scale data format paper from Exeter for a presentation, subject to review of the paper	
2.	Follow up on discussion points about VCF/BCF indexing improvements and consider revisiting development of improved index and compression methods (e.g., BGZF replacement by BGZstd) to tackle large file challenges	
3.	Continue monitoring and gathering feedback on the Excel spreadsheet use cases to inform the opinion paper and future work of the Future of VCF working group	

Meeting recording:
https://us02web.zoom.us/rec/share/7rEDlwNwbxvXap2iyMinB_R21sUyT75BExn89-77H1_fI3ImX2zvluOvwOvSgMev.tXBuloSg04xNjAAE

Meeting minutes:

1. Pre-meeting discussion: imputation service integration

Before the formal meeting, attendees discussed integrating an imputation service into a meta-imputation analysis workflow.

Key points:

- There is no concise implementation guide beyond the published paper and associated code.
- The proposal is to implement this capability within approximately the next 18 months.
- Before committing engineering effort, the team wants a rough estimate of implementation complexity.
- A coordinated experiment involving multiple reference panels (potentially including TOPMed) is envisioned.
- Albert's group is expected to be contacted in the coming months to discuss the technical details further.

---

2. Review of previous action items

The group reviewed three outstanding actions from the previous meeting:

- Invite the University of Exeter/biobank-scale data format team to present (following review of their paper).

- Continue investigating improvements to VCF/BCF indexing and compression.
- Continue gathering feedback through the landscape analysis spreadsheet.

Little concrete progress had been made, although indexing improvements remain an active area of interest.

---

3. Compression, indexing and data formats

The group discussed emerging compression technologies.

Highlights included:

- Evaluation of **OpenZL**, a new BSD-licensed compression framework developed by the creator of Zstandard.
- Initial testing showed encouraging performance on BAM files.
- BCF testing has not yet been completed.
- The library allows data formats to be described declaratively, enabling automatic optimisation of compression strategies.
- Participants noted this could eventually provide a simpler alternative to existing generic compression approaches (such as those used with Tabix), although it remains an early-stage technology.
- Existing index improvements also remain under investigation, with one contributor noting that current performance issues largely stem from parsing large numbers of index entries.

---

4. Variant normalization and graph genomes

The group discussed future challenges around graph genome representations.

Discussion included:

- Difficulties defining "normalization" when there is no single linear reference.
- Distinguishing coordinate normalization from graph canonicalization.
- Whether future normalization recommendations should explicitly consider graph-based references.

No solution was proposed, but there was broad agreement that future work should take graph genomes into account.

---

5. Blog post

Progress on the second working group blog post has been limited because of competing priorities.

The discussion covered:

- increasing emphasis on practical use cases;
- avoiding unnecessary duplication with previous blog posts;
- aiming for publication before the GA4GH Plenary;
- the GA4GH communications review process required before publication.

The group agreed that even a relatively short blog post would be worthwhile if it could be completed in time.

---

6. GA4GH Plenary planning

The group discussed opportunities for the upcoming GA4GH Plenary in Singapore.

Topics included:

- using the completed blog post as the basis for a Connect session;
- obtaining wider community feedback through that session;
- aligning with the published plenary themes (AI in genomics, healthcare implementation, newborn screening and population-scale genomics).

The preferred approach is to have the blog post available well before the meeting to support any related activities.

---

7. Future community engagement

The group discussed ways to keep the working group active between meetings, including:

- inviting guest speakers working on genomic file formats;
- learning from external projects that may influence future VCF development;
- continuing landscape analysis and monitoring new technologies.

2026-03-17 : Blogpost

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)

Global Alliance for Genomics & Health

Attendees “Name (Affiliation)”:

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Blog post ga4gh_future_of_vcf_draft_blog_post_2026.docx	Ole
4.	A.O.B Feedback received:: We have also received a couple of pieces of feedback on the two use cases that were circulated for review.	

Meeting recording:
https://us02web.zoom.us/rec/play/nkjIODJ33LY8CI7JeWqT32DYqWM4KET478CKHC9BBZhdhWeHIc335gDnZkf4e1i7cTLLrmyVTXuy2ORd.ST1BUtd0VvWS6vRN?eagerLoadZvaPages=&accessLevel=meeting&canPlayFromShare=true&from=share_recording_detail&continueMode=true&oldStyle=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2FwD5VYBCIpVfPUK26xTxYutXtddSLw_ZlDH-0CHwkkPh8p711nbDnPmXkDHwIjJM3.mmoy0YGtbcyXhgOg

Meeting minutes:

00:09:33 Ole: GLnexus paper mentioning pVCF: https://www.biorxiv.org/content/10.1101/343970v1.full.pdf
00:10:08 Albert Smith: I think meant to differentiate from gvcf
00:17:32 James Bonfield: https://github.com/samtools/hts-specs/pull/758
00:18:42 Robert Davies: https://samtools.github.io/hts-specs/VCFv4.5.pdf#subsection.7.1
00:19:59 Robert Davies: https://samtools.github.io/hts-specs/VCFv4.5.pdf#subsection.5.5
00:32:48 James Bonfield: https://github.com/srynobio/vmccl
00:33:10 James Bonfield: https://pmc.ncbi.nlm.nih.gov/articles/PMC7153148/
00:36:56 Robert Davies: VRS normalisation https://vrs.ga4gh.org/en/stable/conventions/normalization.html#normalization
00:59:28 Ole: https://pmc.ncbi.nlm.nih.gov/articles/PMC12802598/ maybe we could

Summary:

- Merging Multi-Sample VCFs: Concerns about file size increase during merging; PVCF metadata needs improvement for effective merging without reprocessing.
- Incremental Sample Addition: Key feature needed to add new samples without full recomputation to avoid IO overhead.
- Variant Normalization Tool Shift: Recommended moving from VT to BCFtools for better support and maintenance of tools.
- Annotation Challenges: VCF lacks controlled vocabularies, impacting clarity and interoperability in annotations across tools.
- Indexing Performance Improvements: Discussions on enhancing index resolution and transitioning to better compression methods for faster access.
- Community Engagement Plans: Documenting requirements for opinion papers and enhancing group visibility at events to attract participation.

VCF Format Challenges and Use Cases

The group focused on key technical challenges around merging multi-sample VCFs and improving variant representation for large cohorts.

- Merging Multi-Sample VCFs Limits Explored
    - Oliver highlighted difficulties merging project VCFs (PVCFs) without tripling file size.
    - PVCF is akin to a multi-sample VCF capturing all variants in a sequencing project.
    - Participants debated whether PVCFs hold enough data to merge incrementally without reprocessing all samples.
    - The principle is to avoid complete reprocessing when adding samples, but current format/tooling may not fully support this.
- Handling Home Reference Blocks and File Size
    - GVCF files use hom-ref blocks to compress non-variant regions, but adding unrelated samples breaks these blocks, inflating file size.
    - Multi-sample VCFs with hom-ref info are feasible for small sets (like trios) but become impractical for large cohorts.
    - Recent VCF 4.5 updates attempt to address scaling issues via local alleles and format extensions like LEN for sample-specific alleles.
    - However, it's unclear if these changes fully solve the file size explosion problem.
- Separation of Format and Tooling Concerns
    - Participants acknowledged some merging issues may stem from tooling limits rather than format restrictions.

    - Format must inherently support necessary info (e.g., ref blocks), or tooling cannot compensate.
    - The discussion emphasized the need for format design to enable efficient incremental merging without full recomputation.
- Variant Normalization and Standardization
    - David Lawrence recommended shifting from VT to BCFtools for variant normalization due to VT's lack of updates since 2018.
    - Oliver and others supported a unified normalization tool covering SNVs and structural variants.
    - Thomas raised complex issues around overlapping alleles and ploidy > 2 annotations needing clearer standard representation.
    - The group recognized gaps in the current VCF spec for allele-level and overlapping variant annotation consistency.

VCF Specification and Tooling Gaps

The meeting revealed ongoing issues with VCF specification completeness, annotation control, and indexing efficiency impacting usability.

- Lack of Controlled Vocabulary and Annotation Standards
    - Sam's original format had namespaces for tags, but VCF lacks a clear controlled vocabulary for INFO and FORMAT fields.
    - Hundreds of user-defined tags exist with no canonical list, risking inconsistent annotation naming and interpretation.
    - The group agreed a canonical list of standard tags would reduce fragmentation and improve tool interoperability.
    - Annotation at allele-level versus site-level remains an unresolved specification challenge.
- Indexing and Performance Bottlenecks
    - Current VCF/BCF indexing uses coarse granularity (~4,000 genomic positions per index point), causing slow lookups on large files.
    - File size can be huge, with index points spanning large file offsets, worsening performance for big datasets.
    - Proposals include creating a new, finer-grained index format possibly integrated into the compressed file.
    - A BGZSTD compression-based indexing approach was discussed, potentially supporting random access by both file and genomic positions.
- Compression and Format Evolution
    - Zlib compression limits deduplication to 32 KB windows, inadequate for large multi-sample VCFs.

    - BGZSTD offers better compression and supports more flexible indexing strategies.
    - Moving index inside the compressed file benefits usage in cloud object stores and avoids sync issues between data and index files.
    - These technical improvements aim to reduce I/O load and speed up variant querying on large-scale datasets.
- Alignment with SAM/BAM Headers and Reference Naming
    - Harmonizing reference sequence naming between VCF and BAM/SAM formats is essential for reliable cross-format usage.
    - Sequence collection support must be consistent across formats to prevent data misalignment.
    - This alignment underpins accurate variant referencing and downstream analysis workflows.

Future of VCF Promotion and Community Engagement

The group plans to raise awareness and gather support via upcoming community events and communications.

- Connect Event Representation Strategy
    - Group member volunteered to represent all future VCF and subgroups at the Connect opening session with a 3-minute overview.
    - A script and project/use case printouts will be prepared to clearly communicate ongoing work and challenges.
    - The booth at Connect will host detailed discussions and encourage broader community engagement.
- Call for Feedback on Challenges and Support Needs
    - The representative requested input on current challenges and specific support needs to convey at the event.
    - This feedback will help focus messaging and stimulate collaboration or resource allocation.
- Draft Opinion Piece in Progress
    - One participant shared a short draft opinion paper summarizing key working group ideas.
    - The draft will be circulated via email for review and further discussion outside the meeting.
    - The opinion piece aims to distill technical requirements and promote the working group’s vision.
- Encouragement for Continued Participation

    - Meeting emphasized the importance of increasing active participation beyond note takers and current contributors.
    - Outreach efforts through events and publications are critical to build momentum for future VCF improvements.

Emerging Alternative Formats and Research

New data formats and research directions were briefly discussed as potential complements or alternatives to VCF.

- Annotated Genomic Data Structure Paper Noted
    - A recent paper on a new format for bio-bank scale data was identified, authored by UK-based researchers.
    - The group considered inviting the authors for a seminar to explore this format's potential.
    - The paper references PVCFs and might offer solutions relevant to ongoing challenges.
- Variation Model Collaboration (VMC) Status
    - VMC was discussed but found to have limited recent activity and some defunct components like test suites.
    - The group noted the need to monitor such efforts but remain cautious about adopting formats without active maintenance.
- Caution Against Premature Toolchain Endorsement

  Participants agreed on focusing on conceptual requirements rather than prematurely backing specific tools or formats.

    - This ensures flexibility and avoids locking into solutions that may not scale or be supported long term.
- Interest in Further Exploration (55:30)
    - The group showed openness to exploring new file formats and compression/indexing strategies that could address current VCF limitations.
    - Further reading and external expert engagement are planned before making any formal decisions.

2026-01-20 : Previous action items

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees: Rob Davies, James Bonfield, Reggan Thomas

Global Alliance
for Genomics & Health

Meeting recording:
https://us02web.zoom.us/rec/play/3KytP8-neoT38_wpkMfPG13nAypEH3Sysp1tRBJ4X66Sf1Cudx3lwYl_iUc99sIftSn7yYBxhZxBSS0.RZ9Nqo4e7foVRsHw?eagerLoadZvaPages=&accessLevel=meeting&canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fus02web.zoom.us%2Frec%2Fshare%2F-9QA-Bmhag8PT40O7o6AYw48tYOVA0gvvsLwn5CYQD6MgZWfRrQp691nbsuP2YSS.xPrdu-w7tpu9wv_6
Meeting minutes:
00:08:28 Reggan Thomas:
https://docs.google.com/document/d/1iNyr_Hg2WA4uxWU8RyhIfFEtWZAhUq5j/edit
00:11:19 Ole: vcf-zarr
00:11:56 James Bonfield:
https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf049/8154315
00:13:12 Reggan Thomas: VCF Zarr presentation by Jerome - presentation
00:32:19 Robert Davies:
https://www.biorxiv.org/content/10.1101/2022.12.22.521488v1.full
00:37:24 Robert Davies: https://www.biostars.org/p/9566003/
00:44:10 Reggan Thomas: 18th Feb
00:44:17 Ole: Reacted to "18th Feb" with 👍
00:45:27 Ole: https://www.ga4gh.org/news_item/scaling-vcf-for-a-genomic-revolution/

Summary:
The group reviewed the VCF format, noting its well-designed chunking and compression, and discussed the need for further comparison with existing approaches. They also considered reaching out to tool builders like Zarr and Savvy for input. Ole Schulz proposed writing a blog post to summarize the group's discussions and findings, aiming for a draft by summer. The next meeting is scheduled in three weeks.

Action Items

- Collect bullet points and ideas for a blog post summarizing the group's discussions, candidate file formats and tools, and proposed outreach to format/tool developers, and iterate toward a draft for community contribution (aiming for a draft by this summer).
- Resume investigations comparing variant formats (including VCF) and produce a comparison deliverable for the group to review, providing initial results in about three weeks.

Review of Call to Action and Document Sharing

- •Reggan Thomas confirms that a comms request has been created to send the call to action to the wider audience.•Ole Schulz requests Reggan to share the draft

message in the chat.•Reggan Thomas posts the document in the chat, and an unknown speaker asks for access to the document.•Reggan Thomas grants access, and the unknown speaker confirms it works.

Discussion on Use Cases and Format Comparisons

- •Ole Schulz mentions the need to revisit use cases and an Excel sheet, seeking input from the group.•Speaker 1 recalls a task from months ago to compare different formats, mentioning a talk on VCF.•Ole Schulz and James Bonfield discuss the VCF format, with James noting its well-designed chunking and compression.•Ole Schulz and James agree that the VCF format has good ideas but needs further comparison with existing approaches.

Feedback on VCF Format and Task Assignments

- Ole asks if anyone was impressed by the VCF talk and mentions action items from the meeting.•Ole Schulz shares a link to the VCF paper and asks if anyone wants to pick it up.•James Bonfield expresses that the VCF format seems well-designed but hasn't looked at it in detail.•Ole Schulz and James discuss the need for further investigation and comparison with other formats.

Call to Action Message and User Stories

- Albert inquires about the call to action message and whether it needs further refinement.•Reggan Thomas explains the process of getting the message reviewed and posted to the wider audience.•Reggan Thomas and Speaker 2 discuss the number of user stories to include in the document.•Reggan Thomas suggests focusing on two user stories for now to increase engagement.

Engagement and Future Directions for the Group

- •Ole Schulz questions the group's future direction and the need for more engagement.•Daniel suggests taking small steps and focusing on baby steps to achieve progress.•James mentions the need to reach out to tool builders like Tsar and Savvy for input.•James Bonfield highlights the potential of reaching out to current tool builders for valuable input.

Exploring Noodles Library and Potential Collaborations

- James Bonfield introduces the Noodles library, which implements readers and writers for various file formats.•Ole Schulz and James discuss the potential of inviting Michael Vaka, the author of Noodles, for a talk.•James Bonfield suggests that it might be premature to involve Michael until the landscape analysis is complete.•Rob mentions the Biostars page and a wish list from Daniel Cameron, including a haplotype-centric format.

Proposal for a Blog Post and Action Items

- •Ole Schulz proposes writing a blog post to distill the group's discussions and make them accessible to the community.•James Bonfield supports the idea of a blog post and suggests inviting guest contributors from VCF and Savvy.•Ole Schulz takes the action item to start writing bullet points and ideas for the blog post.•The group agrees to iterate over the next few meetings to finalize the draft and reach out to potential contributors.

Next Steps and Meeting Conclusion

- Ole Schulz asks Tim about the timeline for his tasks, and Tim suggests three weeks.•Tim requests a clear explanation of the group's mandate and tasks, which Ole Schulz provides.•Tim confirms understanding and mentions having something to contribute in three weeks.•The meeting concludes with everyone expressing their availability and readiness to continue working on the discussed tasks.

2025-11-25 : Review landscape analysis

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees “Name (Affiliation)”:

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Plans for 2026	
4.	A.O.B Next meeting on Dec 23 at 2 pm GMT - Cancel?	

Global Alliance
for Genomics & Health

Meeting recording:
https://us02web.zoom.us/rec/share/z2r8RNkREIYbEJP3YgOfiBPHg0r9s0XTXU7F5I4UY4KC4CNc5oM_ag062udnlWTW.AigXE_S-jEEgxEhE

Meeting minutes:
· Explore the use of Zar for storing and processing large genomic datasets, such as the All of Us dataset.
· Investigate the reasons why the Zar format may be larger than the VCF format for certain datasets, as observed in the All of Us data.
· Follow up with Conrad Karczewski to further understand the compression and storage differences between VCF and Zar formats.

Review working_doc-Future_of_VCF_Landscape_Analysis-211006	Tim Hefferon	Next meeting
Update the existing VCF format comparison spreadsheet to include any new contenders mentioned in the Biostars thread	Group	Next meeting
Investigate the pan-genome data representation formats and see if any of them should be added to the VCF format comparison	Group	Next meeting
Attend the upcoming Connect session on the Hail team's sparse VCF format and provide feedback to the group	Group	Next meeting

---

2025-09-30 : VCF Zarr

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees “Name (Affiliation)”: Jerome Kelleher Rob Davies,Andrew Lee, Tim Hefferon, Lee Lichtenstein, Daniel Cameron

Apologies: James Bonfield

	Actions Arising	Assigned To	Deadline
1.	Explore the use of Zar for storing and processing large genomic datasets, such as the All of Us dataset.		
2.	Investigate the reasons why the Zar format may be larger than the VCF format for certain datasets, as observed in the All of Us data.		

3.	Follow up with Conrad Karczewski to further understand the compression and storage differences between VCF and Zar formats.		

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	Ole / Albert
2.	Previous Actions Review	Ole / Albert
3.	VCF Zarr format - presentation	Jerome Kell…
4.	A.O.B Connect - In-person is closed, but virtual is still open and free - https://broadinstitute.swoogo.com/ga4gh13plenary/7101555?i=86R5yrXkzEXnYYCMXah4E96u8i26-_yG Please register to access the zoom link to join the sessions. - Next meeting : 28 October at 9 pm BST / 5 pm ET / 2 pm PT	

Meeting recording:
https://us02web.zoom.us/rec/share/a2GWH3pEx7ak2EyjsVrzSHjlR3fTIgfpMSogwmC5SUi2tOz4TIUpJrYUnmdWIegU.GLb9p_2T2kzLM5tI

Meeting minutes:
RD - Did you use libdeflate for your bcftools tests?
Summary :
Jerome discussed VCFZar, a columnar format for storing high-dimensional scientific data, particularly useful for large-scale genomics. VCFsar offers efficient access by variant and row, reducing storage space significantly. For example, the genomics England AG v2 data set saw a five-fold reduction in size. Jerome highlighted the benefits of using cloud object stores like S3 for scalable, high-performance computations, achieving up to 47 GB/s decode performance. He emphasized VCFsar's extensibility and compatibility with existing tools, suggesting it as a potential cloud VCF standard.

Action Items

- Explore the use of Zar for storing and processing large genomic datasets, such as the All of Us dataset.
- Investigate the reasons why the Zar format may be larger than the VCF format for certain datasets, as observed in the All of Us data.
- Follow up with Conrad Karczewski to further understand the compression and storage differences between VCF and Zar formats.

Jerome's Introduction and Background

- Jerome introduces himself as a group leader at the Big Data Institute in Oxford, working on large-scale genomics data.
- He explains his work on inferring recombinant haplotypes from large data sets and the need for efficient data encoding.
- Jerome mentions his slides and the summary of a paper published in Giga Science.
- He notes that the slides are a subset of a talk he gave in Cambridge and assumes the audience is familiar with most of the content.

Overview of Zarr Format

- Jerome explains that Zarr is a columnar format for storing high-dimensional scientific data, used by major scientific data sets.
- He highlights the key insight behind Zarr: splitting data by calling and storing individual arrays as simple typed binary data.
- Zarr is gaining traction in biological applications, particularly in spatial omics and imaging.
- Jerome mentions the various implementations of Zarr in different languages and the software support available.

VCFs and Zarr Implementation

- Jerome discusses the mapping of the VCF data model into the Zarr format, creating a simple specification.
- He introduces bio desire, a tool that converts VCF data to Zarr at scale.
- Jerome explains the client-side software that emulates VCF tools and the SG kit package that works natively with Zarr.
- He demonstrates how Zarr splits monolithic VCF data into smaller, more manageable chunks.

Data Storage and Access in Zarr

- Jerome describes how Zarr stores data by field and within fields, breaking data into a regular grid of chunks.
- He explains the mapping of chunks to keys in the storage format and the flexibility of storing chunks in various ways.
- Jerome details the process of decompressing and storing typed data in Zarr.
- He emphasizes the efficiency of accessing data by variant and row in Zarr compared to VCF.

Benchmarks and Real Data Examples

- Jerome presents benchmarks using simulations to demonstrate the efficiency of Zarr in storing and processing variant data.
- He compares the storage space required by different formats, including plink, gzip, VCF, and BCF.
- Jerome highlights the competitive compression and processing speed of Zarr.

- He discusses real data examples, including genomics England AG v2 data, chip array data from future health, and all of us data.

Challenges and Future Directions

- Jerome addresses the challenges of dealing with multi-allelic variations and structural variants in Zarr.
- He explains the concept of ploidy and how it affects the encoding of genotypes.
- Jerome discusses the potential for Zarr to serve as a cloud VCF standard, offering interoperability and efficiency.
- He mentions the extensibility of Zarr to include additional metadata and dimensions for different types of data.

Questions and Discussions

- Albert asks about the suitability of Zarr for file transfer and the procedure for updating Zarr data.
- Jerome explains that Zarr is best suited for data sets that need to stay in situ on the cloud and for direct computation on a single shared copy.
- He discusses the efficiency of adding sample chunks to Zarr data and the challenges of dealing with large data sets.
- Jerome answers questions about the encoding of multi-allelic variations and the inclusion of phasing information in Zarr.

Conclusion and Final Remarks

- Jerome concludes his presentation and opens the floor for questions.
- Lee L asks about the potential use of Zarr in the all of us environment and the challenges of using dense formats.
- Jerome explains the compression scheme used in VCF and the differences in storage requirements between VCF and BCF.
- The meeting concludes with Jerome offering to answer any further questions and thanking the participants.

2025-06-10 : Landscape analysis

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees: Daniel C, Andrew Lee, Dmitry Repchevsky

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	Ole / Albert

2.	Previous Actions Review	Ole / Albert
3.	Review the Excel sheet with use cases	Ole / Albert
4	Review of GA4GH Connect session in April (“federated gnomAD”)	Ole / Albert
5	Analysis ready VCF at Biobank-scale using ZARR https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaf049/8154315	Ole / Albert

	Previous Actions	Assigned To	Deadline
1.	Compile a white paper	all	
2.	Review working_doc-Future _of_VCF_Landscape_Analysis-211006	Tim Hefferon	Next meeting
3.	Update the existing VCF format comparison spreadsheet to include any new contenders mentioned in the Biostars thread	Group	Next meeting
4.	Investigate the pan-genome data representation formats and see if any of them should be added to the VCF format comparison	Group	Next meeting
5.	Attend the upcoming Connect session on the Hail team's sparse VDS format and provide feedback to the group	Group	Next meeting

---

2025-03-18 : Gnome AD

Chair: Albert Smith, Ole Schulz-Trieglaff (Illumina)
Attendees:

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	A.O.B 1) April Connect, Heidi Rehm has proposed a session on **Federated gnomAD and the Future of VCF**, scheduled for **Tuesday, April 1, 2025, from 14:15 to 15:45 ET** **Session description**: Multiple groups around the world are working to process their genomic data for sharing in gnomAD. These groups are dealing with many facets of	Albert / Ole

	large-scale genomic data processing that could benefit from the input of GA4GH. For example, we have implemented VRS in gnomAD and will be encouraging other sites to use as well. Also, gnomAD has proposed a new spec to GA4GH to scale genomic datasets sub-linearly. These innovations have been back-ported to the VCF spec and have been are proposed to the GA4GH Future of VCF Working Group as a future revision to the VCF standard. This new format translated into a 50-fold reduction in the size of gnomAD v4 from an estimated ~900TB for a VCF to ~18TB for the sparse VDS, the Hail representation of the same genomic data. The gnomAD and Hail teams hope to engage the GA4GH Future of VCF Working Group, with input from Federated gnomAD sites and the broader community around this new sparse VDS format. Please add the agenda items here as early as possible. 2) Surveying other approaches	

	Previous Actions	Assigned To	Deadline
1.	Compile a white paper	all	
2.	Review working_doc-Future _of_VCF_Landscape_Analysis-211006	Tim Hefferon	Next meeting
3.	Update the existing VCF format comparison spreadsheet to include any new contenders mentioned in the Biostars thread	Group	Next meeting
4.	Investigate the pan-genome data representation formats and see if any of them should be added to the VCF format comparison	Group	Next meeting
5.	Attend the upcoming Connect session on the Hail team's sparse VDS format and provide feedback to the group	Group	Next meeting

Meeting minutes:

Quick walkthrough of the slide deck prepared by Rob on VCF Timeline

Rob and Albert will be joining for the GnomAD/VCF session virtually.

The meeting discussed the preparation for a presentation at Connect, focusing on VCF versions and local alleles. Rob mentioned a timeline of VCF versions and the challenges of handling large datasets. Albert highlighted the **lack of federation in data collection** and the need for better data combination. The conversation also covered the **importance of space reduction in pipelines**, the use of parquet for better compression, and the challenges of managing large datasets.

Actions:

- Explore more efficient data compression and storage techniques for handling large genomic datasets.
- Coordinate with Jonathan to include a segment in the presentation.
- Prepare a presentation on the timeline of VCF version changes for the Connect event.

Feedback on Connect Presentation and Participation

Albert asks Rob for thoughts on the Connect presentation and mentions the enthusiasm of the participants.
Rob shares that a presentation for Connect has been started, focusing on the timeline of VCF versions.
Albert mentions Jonathan Lefebvre's involvement in the presentation and the broad nature of the conversation.
Albert discusses the possibility of attending the Connect session in person or in spirit.

Clarification on Federated Nomad and Data Processing

Albert seeks clarification on the concept of federated Nomad and its implications.
Rob explains that the data collection is not truly federated as it is centralized.
Albert recalls past issues with combining data from different barrier browsers and mentions previous discussions with Nomad.
Rob confirms the lack of federation in the current setup and expresses enthusiasm for learning about scaling solutions.

Discussion on Local Alleles and VCF Changes
Speaker 3 inquires about additional savings methods mentioned in the conversation.
Rob explains the inclusion of local alleles and wrapping boxes as significant cost-saving measures.
Rob outlines the presentation plan, focusing on historical changes in VCF versions and the challenges of local alleles.
Albert asks for input on the presentation, and Rob confirms they are on track to finish before the meeting.

Review of VCF Version Changes and Compression Techniques

Rob presents a preview of the VCF version changes, highlighting significant updates from VCF 4.1 to 4.5.
Albert expresses curiosity about the changes and their impact on HTS GDK development and HDS JDK.
Rob details the additions and modifications in each version, including symbolic alt alleles and structural variants.

Albert and Rob discuss the importance of compression techniques and the use of formats like parquet for better efficiency.

Final Remarks and Next Meeting Details
Albert reviews the meeting agenda and notes the absence of specific topics for discussion.
James Bonfield discusses the importance of space reduction in pipelines and the need for more aggressive data management.
Albert mentions the practice of sharing stripped-down genotypes for PubMed work and the challenges of managing large datasets.

2025-02-18 : Roadmap 2025

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees “Name (Affiliation)”: James Bonfield ,Rob Davies, Tim Hefferon,Reggan Thomas

Apologies:

	Actions Arising	Assigned To	Deadline
1.	Compile a white paper	all	
2.	Review working_doc-Future _of_VCF_Landscape_Analysis-211006	Tim Hefferon	Next meeting
3.	Update the existing VCF format comparison spreadsheet to include any new contenders mentioned in the Biostars thread	Group	Next meeting
4.	Investigate the pan-genome data representation formats and see if any of them should be added to the VCF format comparison	Group	Next meeting
5.	Attend the upcoming Connect session on the Hail team's sparse VDS format and provide feedback to the group	Group	Next meeting

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Roadmapping -2025	
4.	A.O.B	

Global Alliance
for Genomics & Health

	Next meeting - Tuesday 18,2025 at 9 pm BST Register for connect - https://broadinstitute.swoogo.com/connect25bos April Connect, Heidi Rehm has proposed a session on **Federated gnomAD and the Future of VCF**, scheduled for **Tuesday, April 1, 2025, from 14:15 to 15:45 ET** **Session description**: Multiple groups around the world are working to process their genomic data for sharing in gnomAD. These groups are dealing with many facets of large-scale genomic data processing that could benefit from the input of GA4GH. For example, we have implemented VRS in gnomAD and will be encouraging other sites to use as well. Also, gnomAD has proposed a new spec to GA4GH to scale genomic datasets sub-linearly. These innovations have been back-ported to the VCF spec and have been are proposed to the GA4GH Future of VCF Working Group as a future revision to the VCF standard. This new format translated into a 50-fold reduction in the size of gnomAD v4 from an estimated ~900TB for a VCF to ~18TB for the sparse VDS, the Hail representation of the same genomic data. The gnomAD and Hail teams hope to engage the GA4GH Future of VCF Working Group, with input from Federated gnomAD sites and the broader community around this new sparse VDS format.	

Meeting minutes:
Ole - https://hail.is/docs/0.2/vds/index.html

JB https://www.biostars.org/p/9566003/

JB https://github.com/pangenome/pgvf-spec

The list of file output formats from the Human Pangenome Reference Consortium is also vast!

https://github.com/human-pangenomics/hpp_pangenome_resources
They don't seem to have a VCF alternative, unless it's under the guise of the all-singing all-dancing "vg". Conceptually the graph is a variant call, but it's the first 8(?) columns - missing the FORMAT fields and per-sample elements.

Actions:

- Update the existing VCF format comparison spreadsheet to include any new contenders mentioned in the Biostars thread
- Investigate the pan-genome data representation formats and see if any of them should be added to the VCF format comparison
- Attend the upcoming Connect session on the Hail team's sparse VDS format and provide feedback to the group
- Spreadsheet to be be updated - Tim Hefferon

Review of Previous Meeting Action Items

- Reggan shares his screen to review the action items from the October 2024 meeting
- Action items include working on a challenge proposal, reviewing the VCF landscape analysis document, and reaching out to UK pay bank
- Discussion on writing a blog post to highlight the benefits of VCF 4.4 challenges and exploring the current capabilities of GA4GH
- Mention of data compatibility issues with VCF files and the availability of dog genome data for testing scalability challenges

Future Directions and Connect Session

- Discussion on prioritizing the support of current VCF specifications rather than developing new versions.
- Mention of a Connect session planned for April 1st, focusing on federated genome ad and the future of VCF.
- Explanation of the Connect session being in-person with virtual participation available in Boston.
- Tim questions the resources and time available for the group to dedicate to the topic.

Challenges and Resources for VCF Scalability

- Discussion on the importance of VCF scalability and the limited time and resources available.
- Tim mentions a big project after next month that might free up more time.
- James asks about the resources and engagement for the Connect session.
- Discussion on the need for better binary representations of VCF and the challenges of implementing new standards.

Pan Genome and VCF Compatibility

- Discussion on the challenges of representing pan genome data in VCF format.
- Mention of the GFA format and its restrictions on overlapping nodes.
- Explanation of the need for a linear reference genome in VCF.
- Discussion on the potential for new formats to support pan genome capabilities.

Exploring Alternative Formats and Future Standards

- Discussion on the various VCF alternatives and their compatibility with existing standards.
- Mention of AWS omics and its non-file format nature.
- Discussion on the need for a replacement for VCF that supports pan genome capabilities.
- Mention of the Pan Genome Reference Consortium and their use of different formats.

Updating the Excel Table and Future Meetings

- Discussion on updating the Excel table of variant storage options.
- Tim volunteers to update the table and have it ready for the next meeting.
- Mention of the next meeting on March 18th and the need to review the updated table.

- Discussion on the importance of having a comprehensive and up-to-date table for informed decision-making.

2025-01-21 : Cancelled due to unavailability of the leads

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees “Name (Affiliation)”:

Apologies:

	Actions Arising	Assigned To	Deadline
1.			
2.			

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

	Previous Actions	Assigned To	Deadline

Meeting recording:
Meeting minutes:

2024-11-26 : Connect planning

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)

Global Alliance for Genomics & Health

Attendees “Name (Affiliation)”:

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	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. 2) Next meeting on Dec 24,2024 at 2 pm BST - Cancel?	

2024-10-29 : Public data set

Chair: Albert Smith, Ole Schulz-Trieglaff (Ilumina)
Attendees “Name (Affiliation)”:James Bonfield (Sanger), Rob Davies (Sanger), Lon Phan NCBI/NLM/NIH, Tim Hefferon.Reggan Thomas (GA4GH)

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	Albert/Ole
2.	Previous Actions Review	
3.	Public data set - 10k dog genomes - https://ngdc.cncb.ac.cn/dog10k/ ? VCFs: https://kiddlabshare.med.umich.edu/dog10K/SNP_and_indel_calls_2021-10-17/	Albert / Ole
4.	how to publicize the new scalability changes in the VCF spec	Albert / Ole

	Previous Actions	Assigned To	Deadline
1.	Work on the challenge proposal	Albert	

2.	Review working_doc-Future _of_VCF_Landscape_Analysi…	all	
3.	Reach out to the UK biobank	Albert	
4.	Come up with a plan on compression	Albert / Ole	

Meeting minutes:

Status of htsjdk - Broad is responsible. Broad undergoing some changes. At fulcrum - thinking about working on htsjdk. Need to understand the requirements.

YF - Scalability changes - Users of the HTS spec and the VCF spec are not generating 10K . Working on small set of sample.
We need to keep the backward compatibility for the small users
OS : Should have a blog post or some opinion paper that just explains a little bit more of the problems that the local alleles try to address. I think there's sort of like information in various places, but I sometimes have the impression what's missing is sort of like a an easy to grasp illustration of what actually the problem.
Histogram that shows the frequency of occurrence
RD : defining the data model for VCF, which was never really been done
TH - Would like to see data model that built that would deal with pan genomes
JB: check whether or not all those use cases actually fit into htsget
 Split into serialization protocol and data model.
data model is basically your much the language you were using for describing your queries,s o you know what you're storing
OS - like the idea of data transfer and knowledge sharing. The number of institutions worldwide that are really running genomics at scale an end that want to share data is relatively small unfortunately.

Summary:

The meeting discussed the status and future direction of the HTS and VCF specifications. Yossi Farjoun highlighted the need for funding to support HTSget development and emphasized the importance of maintaining backward compatibility for small user sets. James Bonfield noted the challenges of scalability and the need for a unified data model. The group considered using dog genome data for testing but acknowledged its limitations. They also discussed the potential for pan-genome data integration and the importance of supporting existing VCF features before advancing to new versions. The meeting concluded with a call for concrete action and potential funding support.

Action items:
Explore the possibility of writing a blog post or opinion paper to explain the benefits of the VCF 4.4 changes.

Review the current capabilities of HTS-get and how it could be used to facilitate data transfer and querying.

Investigate the availability of large, representative datasets (e.g., dog genomes) that could be used for testing and benchmarking.

Agenda and Initial Discussion
Yossi Farjoun explains his interest in remaining engaged with the community and ensuring no rash decisions are made regarding Java support
Albert Smith inquires about the status of HtsJDK development, and Yossi Farjoun mentions the need for grants to understand the needs better.
Yossi Farjoun discusses the hope for one or two-year funding and the need for intelligent grant writing.

Updates on BCF Spec and Scalability

Albert Smith mentions the update to the BCF spec, specifically version 4.4, which includes local alleles for compression advantages.

Albert Smith discusses the scalability changes in VCF and the need for further aspirations and documentation.

Yossi Farjoun expresses concern about the scalability being important for large datasets but notes that most users work with small sets of samples.

Albert Smith and Yossi Farjoun discuss the balance between addressing scalability for large datasets and maintaining compatibility for small users

Concerns About 4.4 and Structural Variation
James Bonfield clarifies that the 4.4 changes were primarily for structural variation and single sample data representation.

Yossi Farjoun acknowledges the importance of structural variation notation in VCF and agrees with James Bonfield.

Albert Smith questions whether the issues of a few users should be prioritized over the needs of many, leading to a discussion on the scalability and maintainability of the spec.

Yossi Farjoun suggests that eventually, users will push for changes to implement the 4.4 features if they find them necessary.

Discussion on Data and Compatibility Issues

Albert Smith mentions issues with VCF files flagged as 4.3 but not using 4.3 features, leading to compatibility problems.

Yossi Farjoun and James Bonfield discuss the importance of maintaining compatibility and the challenges of addressing scalability and structural variation.

Ole Schulz suggests creating a blog post or opinion paper to explain the problems local alleles address and their benefits.

Albert Smith and Yossi Farjoun discuss the target audience for such a document and the need to justify the existence of the spec to users and funding agencies.

Exploring Alternative Data Sources

Albert Smith mentions the availability of dog genome data as a potential alternative for testing scalability issues.

James Bonfield and Rob Davies discuss the limitations of dog genome data and the need for larger, more varied datasets.

The group considers the possibility of using bacterial or viral data but acknowledges the challenges of finding suitable datasets

Albert Smith suggests revisiting previous hackathons and workshops to see if any outcomes could be useful for current discussions.

Future Directions and Action Items

Rob Davies proposes focusing on supporting the current VCF spec rather than developing new versions.
Yossi Farjoun emphasizes the need to support the existing spec before moving forward with new developments.
James Bonfield suggests reviewing the current implementation of Htsget with VCF to ensure it meets all use cases.
The group discusses the need for a data model that can handle pan genomes and integrate structural and small variation data.

Support and Funding Considerations

Yossi Farjoun mentions the possibility of obtaining a letter of support from GA4GH for grant applications

Albert Smith and Yossi Farjoun discuss the importance of having proper support in the backend to use the spec effectively.

The group acknowledges the challenges of funding and time but agrees on the need to focus on supporting the current spec and addressing scalability issues.

2024-10-01: Cancelled (due to unavailability of the leads)

Chair: Albert Smith, Ole Schulz-Trieglaff (Illumina)

	Previous Actions	Assigned To	Deadline
1.	Work on the challenge proposal	Albert	
2.	Review working_doc-Future _of_VCF_Landscape_Analysi…	all	
3.	Reach out to the UK biobank	Albert	
4.	Come up with a plan on compression	Albert / Ole	

2024-09-03: NIH roadmap and VCFcompression challenge

Chair: Albert Smith, Ole Schulz-Trieglaff (Illumina)
Attendees “Name (Affiliation)”: Rob Davies, Daniel Cameron ,James Bonfield,Christopher Chang,Reggan Thomas (GA4GH),Tim Heffreon
Apologies:

	Actions Arising	Assigned To	Deadline
1.	Come up with a plan on compression		
2.			

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	Albert/Ole
2.	Previous Actions Review	
3.	Review of the NIH roadmap	Albert / Ole
4.	VCF compression challenge	Albert / Ole
5.	Next meeting on Oct 1, 2024 at 9pm BST/4 pm ET 1) Connect 16/17 Sep.	

Global Alliance
for Genomics & Health

	What's new in VCF v4.5 Monday, 16 September 2024, 11:15 - 12:45 AEST Agenda File Formats Working Practices Monday, 16 September 2024, 04:15 pm - 5:45 pm Agenda - Yet to be finalised	

	Previous Actions	Assigned To	Deadline	
1.	Work on the challenge proposal	Albert		
2.	Review working_doc-Future _of_VCF_Landscape_Analysi…	all		
3.	Reach out to the UK biobank	Albert		

Meeting recording:
https://us02web.zoom.us/rec/share/hRABxp0IhGU9bg93QDfjXKTPedt2T__iAcno8trwm6pJvf6AHXj0w3mfzmSy2BuL.1SWMBw5Vc4H8KB7-

Meeting minutes:
OS - Any burning issue
AS - on the upcoming connect session - VCF 4.5
AS - Illumina has used any local alleles - Uk biobank?
OS - due to be released. Has it's own policy cycles
OS - UKbiobank - Illumina hasn't used multi sample vcf.
OS - Ole and Albert to connect offline to discuss about htslib

VCF compression challenge -
AS - https://academic.oup.com/gigascience/article/2/1/2047-217X-2-5/2656124
JB - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9205427/
CC - https://academic.oup.com/gigascience/article/6/5/gix009/3057432
AS - Need to come up with a plan on compression
RD - https://www.ga4gh.org/product/variant-benchmarking-tools/
OS - need to come up withdata set that can be shared.
RD - https://ngdc.cncb.ac.cn/dog10k/

Meeting summary:

NIH Roadmap and VCF Compression Challenge

- Ole Schulz initiates the meeting, mentioning the absence of Albert and asking for agenda items.
- Albert Vernon Smith joins late and inquires about the status of a presentation on VCF 4.5.

- Daniel Cameron explains the scalability issues in VCF 4.5 and the stability of certain features.
- Discussion on the implementation of VCF 4.5 features and the need for tool updates.
-

Implementation and Tool Compatibility Issues

- James Bonfield discusses the implementation of local alleles in HtS lib and the need for updates.
- Rob Davies mentions the lack of support for certain VCF features in BCF tools.
- Albert Vernon Smith suggests potential contributions from Illumina to improve HDS lib support.
- Ole Schulz and Albert Vernon Smith discuss the use of HtS lib in UK Biobank analysis and the potential for further contributions.
-

Challenges in Organizing Compression Challenges

- Albert Vernon Smith and James Bonfield discuss past compression challenges organized by Pistoia Alliance and Huawei.
- James Bonfield highlights the challenges of organizing such competitions, including the need for dedicated resources.
- Discussion on the potential benefits of compression challenges and the need for clear evaluation criteria.
- Rob Davies and Daniel Cameron emphasize the importance of simple and objective evaluation metrics.
-

Potential Data Sets and Synthetic Data

- Ole Schulz and Albert Vernon Smith discuss the need for a large, shareable data set for a compression challenge.
  Albert Vernon Smith mentions the 10,000 Dog Genomes Project as a potential data source.
- James Bonfield suggests the possibility of using synthetic data to upscale from existing data sets.
- Discussion on the challenges of simulating realistic data and the potential issues with synthetic data.
-

Next Steps and Action Items

- Albert Vernon Smith proposes educating themselves on how to organize a compression challenge.
- James Bonfield mentions the use of the GA4GH benchmarking suite for evaluation.
- Ole Schulz and Albert Vernon Smith agree to follow up on potential data sources and organize a plan for the challenge.
- Reagan suggests setting up booths for each workstream at the plenary session to promote use cases.

Planning for the Plenary Session

- Reggan proposes using QR codes to direct interested participants to use cases.

- Albert Vernon Smith suggests updating the blog post to advertise the compression challenge.
- Daniel Cameron offers to incorporate information about the challenge into his presentation.
- Discussion on the logistics of the plenary session, including time zones and session responsibilities.

Finalizing the Plan and Follow-Up

- Ole Schulz offers to create a flyer or mini poster to promote the blog post.
- Reggan agrees to copy paste information from the blog post into the template.
- Albert Vernon Smith emphasizes the importance of following up on the ideas discussed and saving the chat for reference.
- The meeting concludes with a plan to continue discussions offline and follow up on the action items.

2024-05-14: Challenge planning & Publication

Chair:Albert Smith, Ole Schulz-Trieglaff

Attendees “Name (Affiliation)”: Andrew Lee, Gigi Van den Hoef, Daniel Cameroon,Robert Davies, James Bonfield, Jonathan LeFaive, Christopher Chang, Tim Hefferon,Lon Phan (NCBI/NLM/NIH)

	Actions Arising	Assigned To	Deadline
1	Review working_doc-Future _of_VCF_Landscape_Analysis-211006	all	Next meeting
2	Reach out to the UK biobank	Albert	

	Agenda Item	Person/Time
1.	Welcome, Introductions, New Faces, & Confirm Agenda Ole Shulz - New subgroup lead	Albert
2.	Previous Actions Review	
3.	Challenge Planning	Albert/Ole
4.	Transfer the knowledge into a publication - Discussion	Albert /Ole
5.	Next meeting on Jun 11, 2024 at 9pm BST/4 pm ET	

	Previous Actions	Assigned To	Deadline
1.	Work on the challenge proposal	Albert	

Meeting recording:
https://us02web.zoom.us/rec/share/IGHcbZVETZ0xQctFjfCGgqtfFceQw6o5iajZwqT8o_-YjITxqAEiaNeN62vCYzt-.M4E2AD7v6oJfvKzA

Meeting minutes:

DC : VCF 4.5 is finalized , but methylation will change
actual finalization of the specs overall needs another 6 week review period there

AS: when you say scalability, is that both the local allele stuff as well as reference box or, or, you know, they were both under consideration.
DC: the initial implementation actually had reference block size and header
DC: There is an issue with VCF generally in that if you have an Indel, which overlaps in Indel. We overlap in Indel, which overlaps in Indel. You kind of need to chain them together and give all this all the appetite sequences there.
AS: what's this the state of like you know of any you know of any implementation?
DC: Think Illumina
RD: It may true treat unknown number types as numbers equals one, actually.
DC: Scalability VCFs that don't write the end field. Cause in 4.5, and is now calculated.
AS: what's the outstanding issue with the methylation?
DC: Field names and conventions on the methylation because We wanna do base modifications and there's lots of different based modifications.

DC: Base modifications to change in 4.5:
https://github.com/samtools/hts-specs/issues/767

AS: Is there a way that we should be thinking about a way to to do an implementation?For a challenge and get it out there
DC: How once this new spec is out, how will people hear about it?
AS : what is the appetite or thoughts about actually creating a manuscript that actually compares methods?

2024-03-19: Challenge planning & Survey review

Chair:Albert Smith,
Attendees “Name (Affiliation)”:Rob Davies, Jonathon L,James B,Ole S,Christopher C,Lee L,Nina Haffer

	Actions Arising	Assigned To	Deadline
1	Work on the challenge proposal	Albert	

	Agenda Item	Person/Time

Global Alliance for Genomics & Health

1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	• Review of survey • Planning challenge	Albert
6.	A.O.B Next meeting on Apr 16, 2024 at 9pm GMT/4 pm ET	

	Previous Actions	Assigned To	Deadline
1.	Draft 4.5	DC	Before next meeting
2.	Questions for survey for DPs/interested challenge participants	MF/AS	Friday / Done

---

2024-02-20: PR 434/435 & Benchmark Challenge

Chair:Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”:Robert Davies, James Bonfield ,Dan King, Christopher Chang,Jonathon L,Daniel C,Reggan Thomas (GA4GH)

	Actions Arising	Assigned To	Deadline
1	Draft 4.5	DC	Before next meeting
2	Questions for survey for DPs/interested challenge participants	MF/AS	Friday

	Agenda Item	Person/Time
1.	Welcome, Introductions, New Faces, & Confirm Agenda	
2.	Previous Actions Review	
3.	Discuss the benchmarking challenge proposal	Mallory/Albert
4.	Continue discussion about PR 434 & 435	Daniel
5.	Announcement - Nomination for a new Co-lead	Mallory/ Albert
6.	A.O.B Next meeting on Mar 19, 2024 at 2pm GMT/9 am ET	

	Previous Actions	Assigned To	Deadline
1	Mallory to connect with Thomas regarding UK biobank data	Mallory	DONE
2	Group to review proposal	all	

Meeting minutes:
MF - High level discussion on the proposal
RD - Competition between sort of people?
AS - competition probably I don't think is the right term. I think it's more to ensure there's at least pairwise comparisons and and it's not necessarily it's not you know, we're trying to just understand and and just having at least pairwise, you know, hopefully your benchmark

JB: we're really talking BCF here as the model of and the format
We have the data in some proprietary format, which is the information that would be in a VCF, but isn't itself a VCF.

JB - Challenges - general problem is that everybody has their own programs, and their own programs have to all.They all have a steep learning curve. What? Parameters used to get them running. If it's Java, how much memory do you want to give it?
MF - storage and compression aspect, Comparing the time comparing the storage compression, comparing how it functions in a particular analysis

PR 434 and 435 -
Unclear information. Details around the 435 ref block checkpointing.
435 feels like a feature to make the VCF Make giant VCF usable to process directly off those VCF themselves.
DC - Get 4.5 draft.

2024-01-30: Benchmarking and PR434

Chair:Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”:Christopher C, Dan King (Broad/Hail), Surakameth (MOPH), Rob Davies (Sanger), Kahlil Lawless (Illumina), Leon Rauschning (CRG) ,Ole (Illumina), Mihai Todor, Daniel C, Jonathan L, Lee Lichtenstein,Reggan Thomas (GA4GH)

	Actions Arising	Assigned To	Deadline
1	Mallory to connect with Thomas regarding UK biobank data	Mallory	DONE
2	Group to review proposal	all	

	Agenda Item	Person/Time

Global Alliance
for Genomics & Health

1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Renew UK Biobank data access application for VCF scaling testing	Albert
4.	VCF 4.5: how mature PR https://github.com/samtools/hts-specs/pull/434 is (any objections?) and the GA4GH Connect VCF session planned	Daniel
5.	Discuss benchmarking challenge proposal	Mallory
	A.O.B. - Connect meeting (April 21 - 24) at Ascona, Switzerland. Hybrid participation available. - Submit your connect meeting proposal here (deadline - Feb 16,2024)	all

Meeting minutes:
AS - Any idea for challenge evaluation with UK Biobank?
RD - engage biobank but not with an analysis platform?
Mallory to connect with Thomas regarding UK biobank
DC - Planning for a session at GA4GH connect . Planning to have a draft VCF 4.5 before connect
DK - I think all we need is the local alleles and the ref blocks.
Ole - we have published a data set as well in that format. We as Illumina.
AS - I believe there were still some outstanding issues and we hadn't fully addressed the rough blocks and there were some concerns in terms of that.
DK - Illumina is MSVCF only uses locally.?
AS - I believe this is the one https://github.com/samtools/hts-specs/issues/705
MF - request feedback 2 week prior to the connect meeting
DC - what is the scalability issue with the current implementation of ref blocks within VCF?
MF - Discussion on dream challenge
Ole - You're proposing that people basically or participants bring their own data and run their tools on their own data.
MF - I guess there's a risk that people just tweak their message to their own favorite data sets and it makes it really hard to compare different approaches.
MF - yes, people would have their own data and we would request they benchmark at least 2 methods. So One could be there, they have, but they would have to choose from the list.
MF - we would have like a list of methods in the challenge and you have. Maybe we say choose at least 3, but the goal is to pick at least a couple. And then within that one data set you compare across because this is another challenge we ran into.
DK - little skeptical of round tripping of VCF though
DK - We've agreed that the VCF is a bad format for large call sets or maybe I've decided that maybe we don't agree on that.
LL - we don't take in VCFWe take in GVCS. Round tripping to another GVCF is not something We would ever build a method for.
MF - then what would the workflow be? What would be what workflow would make sense for the challenge that we're trying to.

LL - GVS kind of like the GVCF 2 joint call set. My objections really would be along the lines of, you know, no round tripping out at GVCF and even if you are gonna do something like that, you really want concordance, not an empty 5.
JL - hear your round trip concern. But I imagine You also have a concern with starting with something that is in a GVCF.
DK - Broadly, yeah. Yeah, something GVCF
RD - what variable are we trying to optimize here?
MF - speed, size
JL - Some key variables and it's basically can you have a a small file size for an exchange format that doesn't take 100 gigs of RAM to serialize
MF - May we strip off round tripping?

2023-11-28: PR 434

Chair:Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”: Lee L,Ole Schulz-Trieglaff (Illumina),Dylan Spalding (CSC, GDI), James Bonfield ,Rob Davies,Jorge Oliveira (BioData)

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Updates on PR 434 (from James): The main stumbling block to it now is a case that was spotted involving empty lists when no local alleles exist. The quoted example includes such a case: LAA:LGT:LAD:LPL :0/0:30:0 That's LAA of "" before the first colon. There was debate, but it looks like that is illegal in BCF as it has no notion of a zero-length list, and the specification text also states LAA is 1 or more alternative alleles. Bcftools just rewrites this case as "." for "missing", but that may be interpreted by others as an unknown possibly. Our conclusion is basically it just needs the VCF maintainer to decide whether this should be legal, and if not whether "." can be defined to be the same thing.	
4.	Ideas for April Connect meeting in Switzerland? Opportunity for collaborative work and engagement with DPs	

Global Alliance
for Genomics & Health

2023-10-31: PR 434, Blog post

Chair:Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”: James Bonfield, John Marshall, Zhuoyi Huang,

	Agenda Item	Person/Time
1.	Welcome, Introductions, **New Faces**, & Confirm Agenda	
2.	Previous Actions Review	
3.	Updates from Ole (Illumina) ● During the last GA4GH call, I had promised to comment in favour of the “local alleles” change. I have done so here: https://github.com/samtools/hts-specs/pull/434 I am interested in seeing this proposal become part of the VCF spec, if there is something else that you need me to support this, please let me know. ● You had also asked me for an example of an msVCF line that breaks htslib due the number of alt alleles. Unfortunately, since we moved to the local alleles flavour of VCF, we have not seen this problem anymore (which does not mean that it will not occur). I think it should be easy to simulate such an example and I could try to help with that if needed.	
4.	Blog post response: ● Overall, the post performed very well, and is GA4GH’s second most viewed post this year!	Mallory
	A.O.B. ● April Connect meeting in Switzerland - opportunity for some collaborative work	

Meeting minutes:
3. Next steps: raise newest comments with Daniel to get change (local alleles merged). PR needs some cleanup. Target v4.4 for changes. Raise at next File Format 14/15th (Mallory or Albert to attend to push for VCF changes)

Global Alliance
for Genomics & Health

2023-10-03 : Plenary and UK Biobank

Chair: Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”: Robert D, James B, Ole Schulz-Trieglaff (Illumina)

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Reflections on GA4GH Connect/Plenary	All / 15’
4.	Next steps with working with Illumina / UK Biobank for reviewing VCF changes?	All / 15’
5.	Interest in hearing about new universal data compression algorithm by Albert Guillen i Fabregas and Mehdi Dabirnia who reached out to James about engaging the GA4GH LSG WS	All / 15’
6.	AOB ● Blog post published: https://www.ga4gh.org/news_item/scaling-vcf-for-a-genomic-revolution/ ● Next meeting: 31 October	

Meeting minutes:

- Reflections on GA4GH Connect/Plenary
  - Topics raised at GA4GH around how VCF will interact with graph genomes / pangenome references (e.g. Human Pangenome Reference Consortium (HPRC), Australian Reference Genome Atlas).
  - Suggestions to get industry support for maintaining (being champions for) VCF specification.
- Comment on memory explosion with large allele numbers (post meeting thought from AVS):
  Though 18,000+ alleles might not be useful in a data quality and/or biological context, a test case may represent a particular memory explosion which could be followed up. It may represent a logical flaw in VCF design which could be addressed. Local Alleles may address, but having a test case could enable testing of other paradigms.

Global Alliance
for Genomics & Health

2023-09-05 : UK Biobank / Illumina Use Case

Chair: Albert S, Mallory F
Attendees "Name (Affiliation)": James B, Rami M, Amreen M, Christopher C, Robert D, K Singh, Zhuoyi H

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review (covered in points 3 and 4 below)	
3.	Blog post set to be published this Friday - thanks all for your input, especially Mike, James, Oliver, & Geraldine! (sneak peak!)	Mallory / 2'
4.	Discussion about use case - from UK Biobank / Illumina - for Future of VCF	Albert / 20'
	A.O.B.	

→ Who to do the review/merge of the local allele PR: 1) Daniel (Cameron) as the only active main VCF maintainer needs to approve and merge; 2) someone from industry (e.g. Illumina)
→ Need a framework to test?
→ Need some (anonymised) data to test

2023-07-10:Local Alleles & Blocksize

Chair: Albert Smith, Mallory Freeberg
Attendees: K singh,Yossi F,Christopher C,Daniel C,John Marshall,Mike L,Jonathon L,Rob Davies, Oliver Hofmann, Tim Hefferon, Reggan Thomas

	Actions Arising	Assigned To	Deadline
1	Oliver to check with Peter Goodhand for a UK Biobank contact, especially for a Champion for moving forward VCF changes?	Oliver Hofm…	
2	Share draft of blog post for input	Mallory	

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Update on NCBI/NIAID Codethon for VCF Files for Population Genomics: Scaling to Millions of Samples	Albert

3.	Review PRs ● Define Local Alleles in VCF to allow for sparser format: https://github.com/samtools/hts-specs/pull/434 ● Add reference blocksize and checkpointing to VCF: https://github.com/samtools/hts-specs/pull/435	All
4	A.O.B ● Session proposals for Connect 19-20 Sept (before Plenary) - By 30 July, 2023 ● Blog post outline shared with GA4GH Comms team and approved to go ahead. Will share a draft with our WG and request input/review. Aim to finalize by 8 August.	

Meeting minutes:
VCF hackathon –
Sensitive to share project info.
MF – meet as the wg?
AS – What is the end product? Summary report. That's not required. Trying to raise awareness. 80 applicant. Approx.. 50 will be involved.

AS – they will provide covid data set. Cloud resource – amazon .
TH - https://www.ncbi.nlm.nih.gov/sars-cov-2/
TH - https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049

PR 434 –
MF – Open for discussion
AS – Yossi is the original submitter.
YF – struggling with a sample on how to put that into vcf. As the sample grows, vcf grows. The idea is to enable the tools to have original GVCF in avcf format.
Mike Lin implemented this. Alternative option. Never had an agreement on what is better. Stagnant until someone does the comparison.
The fields are there in order to increase the compression. The main thing trying to achieve
The current vcf scales . problem when you reach millions of samples.
VCF is a way to go forward for the population wide. Do we accept that bigquery/.. a better format for storing variant samples.
This was written as a strawman proposal.
AS – In some level in your involvement, not much changes. We had a long discussion where we are trying to land. The reason is coming to this PR, we are trying to make concrete steps.

AS – trying to figure out a mechanism to go forward.

JB – General plan – ideally long term solution - Hail? Not a good format. This PR is basically split into 2.
Local allele – more samples - we still get one value in the matrix.

Reference blocks –

Benefits ?
ML – covered the continued propagation of vcf. Starting from 2 different places. Local alleles and reference blocks. Merge or combine gvcf files. Approached from different starting point
Find out a way to compress that. GVCF file and merging them ? Do I write the end position of the ref block. Multi sample , multi vcf data model.

DC – per sample gvcf ? I see the notation and conversion of gvcf.

ML – GVCF as the starting point
YF – gvcf are linear. Test in spec, gvcf should be part of the test.
DC – 2 different gvcf?
DC – merging doesn't really work. If we break that into 2. The shorter number of
-
YF – when is the beginning of that sample. Break it back into gvcf.
DC- 2 gcvf and merge them, do we loss information?
YF – we need to expand the example.
YF – star allele interact with gvcf
YF – rebase and put it in 4.4. not back version to v 4.3
DC – should be 4.4 or 4.5
JB – not in 4.3 , it is not breaking anything. An example would be helpful
ML - is it true , nomad and associated project use this representation
YF – need to check with LB
AS – Data model and preprint server?
ML – Some degree of production experience would be helpful
ML – I think this PR can move forward. Starting point conceptually GVCF .The whole gvcf merging is important. Cohort dataset is produced. GVCF merging for standardization.
JB – local allele make sense. Not an explosion if we get more than 2 alleles.
JB – The first PR make sense. Are you starting from gvcf .. independent of reference blocks.
DC- ref block is a better solution. In terms of the format or scalability . like the ref block approach. Local allele one. Question on adding an extra type. Reinterpret L prefix?
Common prefix on everything or is this minimal set ? thought on local alleles.
JL – The case of starting to merge VCF , observation or evidence. If you are starting from merge vcf
JM – requires a new version number. Is it compatible ? Sound to me, pick up from GT field and all other fields contain values for allele listed in GT than alt. Much bigger changes . Never being particularly. If we chose to put in 4.5
JL – convert them into old interpretations or pre-defined?

JM – shot list of L prefixed field or grow over time?
YF - the list of fields that need specific alleles is not that long. Doesn't seem to have
Overloading of non prefixes .
DC- Is it purely a compression thing? To merge stuff missing values. Lossless round tripping.
YF – local alleles?
DC- conversion between local and expanded ones.
AS – where do we go from here. We need a champion.
YF – We need to contact a large biobank – uk biobank. You must use and we need to update the spec and we need a champion.
AS – we need to reach out and dig deeper. We tried in the past and we need to try again
OH – Heidi or Ewan help for reaching out to the biobank . maybe peter. To check with peter?

2023-06-12: Connect action items & VCF specification

Chair: Albert Smith, Mallory Freeberg
Attendees "Name (Affiliation)": Mallory, Kshitiz Singh (Queen's University), Christopher, Reggan, Yossi Farjoun, Jonathan, Daniel, Robert Davies (Sanger), Tim
Apologies : Albert Smith

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3.	Discussion - Action items from the Connect meeting : key takeaways" and "next action items	Mallory / 30 mins
4	VCF Spec request for technical support to make progress on the VCF spec	
5	A.O.B • Session proposals for Connect 19-20 Sept (before Plenary) - By 30 July,2023	

Meeting minutes:
CC - https://ncbiinsights.ncbi.nlm.nih.gov/event/vcf-for-population-genomics-codeathon/

Global Alliance
for Genomics & Health

2023-03-20: Synthetic Geno/Phenotypes

Chair: Albert Smith, Mallory Freeberg
Attendees “Name (Affiliation)”: James Bonfield Andrei B,Jonathan L, William Rayner,Kshitiz S,Reggan Thomas,Christopher Chang,Tim P

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Synthetic genotypes/phenotypes for 500k individuals	Will Rayner / Andrei Barysenka
3.	GA4GH Connect session agenda	Albert / Mallory
	A.O.B.	

Meeting minutes:

MF - https://www.cineca-project.eu/
WR - Why we developed the dataset is because - we are working on a federated dataset.
One of the use case - Polygenetic risk score
We could federate an analysis model on excel data
There weren't any biobank scale, synthetic dataset we needed to test the system and verify it is working.
Andrei developed data set which is of 500,000 genome wide

AB - synthetic data we were trying to generate at ITG and as part of Cineca project, and as we already mentioned, our motivation was to generate realistic data sets.
AB - data set properties as genetic as data set, containing genotypes and phenotypes
three different broad approaches to generating synthetic genotypes
our data set consists of both genotypes and phenotypes
AB -how we tried to generate genotypes. So, one approach is for time so basically we start with number of individuals
simulate the evolution of those individuals, model selection and population
computationally too heavy because if you are interested in hundreds of 1000s, or millions of individuals, it's, it's gonna take a very long time to generate this kind of synthetic data
AB - we can directly model selection the population structure, but it's not feasible for our purposes. Another simple approach is based on resampling. So basically, we start with some genotype data.

AB - several different approaches to generate synthetic data, one is forward time based approach, which is computationally heavy and the second approach is easily parallelizable. That is, the advantage of this method, but disadvantage is this resampling method does not have this method. If we are trying to generate our database on some real data, the source of chromosome pieces, the source of genetic data can be potentially identified.
AB - decided to use a combination of two methods. Basically, it's a combination of the resampling method and coalescent based method

JL - Are there benefits to using the coalescent approach as a seed for resampling as opposed to resampling a public genotype dataset like 1000 genomes?
AB - probably not a lot of people would object but if we're using some other datasets
AB - reassembly approach based on real data, we can potentially trace the synthetic data back to the origin
KS - How do you decide on the values of alpha, alpha, beta, gamma and delta?
AB- user of this pipeline to set those parameters

2023-02-20: SVCR data model & datasets

Chair:Albert Smith, Mallory Freeberg

Attendees: Rob D, James B, Tim P, Christopher Chang,Jonathon L,Likhitha S,Gisli M, Daniel King,Anastasia

	Agenda Item	Person/Time
1.	Welcome, Introductions & Confirm Agenda	
2.	Previous Actions Review	
3	SVCR data model and encoding in VCF	Tim Poterba
4	Review of available data sets,Synthetic data set	Albert/Mallory
5.	Connect meeting discussion	Albert/Mallory
	A.O.B.	

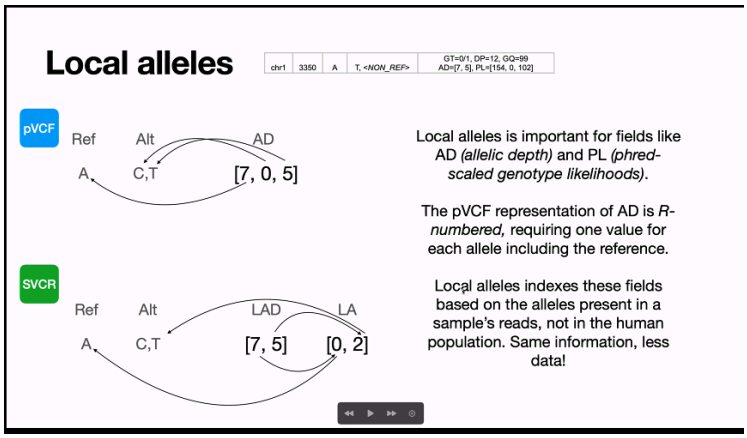
Meeting minutes:

Walkthrough of SVCR model and encoding in VCF - Poterba SVCR GA4GH 2023.02.20.pdf

TP - Representation is an abstract data model. And it can be represented in terms on disk in terms of bytes in a variety of formats
TP - when we say that project VCF doesn't scale well really, the problem there is that the representation, the data model doesn't scale well. And however you choose to represent that

information in a file. Even if you get really good compression, you're still going to be scaling super linearly.

So the problem with with Project VCF is that the data model is that svcr is a data model abstract representation, which we can represent in any structured format, like VCF with some adjustments to the spec, just a couple as I'll describe later, or CSV, JSON, Avro parquet, what have you and I think, for me, at least the most important thing for this group to to agree on is what is our data model for variant calls for large cohorts because once we've done that, I think converting between VCF or any of these other formats is relatively painless compared to the monumental task of recoding 10s of terabytes of data anyway. So spcr At its core is a MultiSample generalization of GVCS



Cc: one other way to make single sample extraction a bit cheaper is to shard a bit by sample subgroup. But if you can just store the original sample gVCFs, that's easier and faster.

JL -Need for LA?

TP - The number of values present in la is larger than pVCF.

GA4GH Connect planning to attend: Mallory, Albert, Rob, James,

---

2023-MM-DD: Template

Chair:
Attendees "Name (Affiliation)":

	Actions Arising	Assigned To	Deadline
1			

2			

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

Further disc



Global Alliance
for Genomics & Health

GA4GH Future of VCF

Minutes and Actions 2021

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

Table of Contents

Meeting Protocols

Agendas, Minutes and Actions
- 2021-12-08: VRS/VCF Alignment
- 2021-11-29: VCF Scalability
- 2021-11-01: VCF Scalability
- 2021-10-04: VCF Scalability
- 2021-08-09: VCF Scalability
- 2021-07-12: VCF Scalability
- 2021-06-30: VCF/VRS Alignment Meeting
- 2015-06-14: VCF Scalability Reboot
- 2021-05-18: Fifth Meeting 2021
- 2021-04-15: Fourth Meeting 2021
- 2021-03-31: VRS/VCF Alignment Meeting Cancelled
- 2021-03-23: Third Meeting 2021
- 2021-02-22: Second Meeting 2021
- 2021-02-03: VRS/VCF Alignment Meeting
- 2021-01-25: First Meeting 2021
- 2020-01-06: Joint VRS/VCF Meeting #4
- 2021-MM-DD: Template

Previous Meeting Minutes

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 All deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd

ussions also on the File Formats Minutes



Global Alliance
for Genomics & Health