GA4GH FEDERATED ANALYSIS PROOF OF CONCEPT

Rodrigo Barnes, Chief Technology Officer, Aridhia
rodrigo.barnes@aridhia.com
July-2015

Genomics – the biggest big data science

In early July 2015 four of the world's leading researchers from within the genomics community published a paper which suggested that "…between 100 million and as many as 2 billion human genomes could be sequenced by 2025, representing four to five orders of magnitude growth in ten years…" to become the biggest of all the big data domains, reaching exabase-scale genomics within the next decade. [1]

A common concern within the genomics community, and one which is shared by the authors of the aforementioned paper, is the availability of sufficient data in any one site to come to any valid scientific conclusion and move beyond that into the application of the science in healthcare delivery. As comparable sequencing and variant calling technologies become available that allow consistent analysis, new models of interaction are required which facilitate the collection of vast amounts of data from multiple sites into secure repositories to enable collaborative analysis on a global scale. At its heart this is about clinical communities pooling their knowledge of relevant variants.

As the sequencing data being created independently by multiple projects across the world (such as the US Precision Medicine Initiative) which aim to map genetic variation grows at an exponential rate, our ability to adequately store, share and analyse this data becomes an increasingly urgent issue, one which requires early and detailed consideration of the infrastructure needed to support future growth in the domain.

Is federated analysis the answer?

Federated analysis describes the ability to share data for distributed analysis without physically sharing it. At its core, it is about making connections between distributed groups and enabling them to study important sample data in a collaborative fashion.

Federated analysis therefore provides an ideal foundation for a globally fragmented and distributed genomics community which stores data in isolated databases. It is seen as supporting partnership models of collaboration, where distributed parties agree in principle to cooperate, but are reluctant or unable to pool resources until a real use case is in place.

While some sharing models rely on pooling data, whether in a private, open or commercial framework, federated models aim to respect important local legal, privacy and consent arrangements by allowing relevant data to remain in local storage, reducing the need for data to travel. Researchers are then able to gain access to a larger 'virtual' data set comprising information - if not data - from multiple sites, upon which analyses can be run simultaneously, whereby increasing research efficiency.

Federated analysis therefore promises researchers access to larger sample sizes, facilitating large-scale data comparison to get better insight and drive improvement. This has been highlighted as important in two distinct health-

[1] *Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015 Jul 7;13(7):e1002195. doi: 10.1371/journal.pbio.1002195 http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195*

related domains: genomics, due to the high cost of data and its low mobility, and health improvement, where partnerships share data in order to improve standards of care.

Key elements of federated analysis

Layered Analysis: Technically, federated analysis includes both the interrogation phase where different sources are polled or queried for data, but also the broader workflow of assembling an analytical outcome. This is because a future, successful model of federated analysis may require the execution of end-user code remotely at each site, in which case the analysis is truly distributed.

Low Costs: The overall cost of initial and continued participation must be clear and low. This is important since it is easy to see how federation might be valuable, but more difficult in practice to get stakeholder approval and funding in place to execute a given analysis. Once prerequisites are in place, a federated analysis system should be able to commence and deliver results quickly.

Privacy Protocols: Interoperability between centres is necessary but not sufficient. Privacy preserving protocols are key to enable local stakeholders to approve participation. Layered protocols allow a user to build up a federated analysis or report incrementally. At each stage a contributing party may check their willingness to participate in a subsequent layer.

Complimentary Technologies: A range of technologies are required to implement federation. At the time of writing, these are not in widespread use, or are even of a production quality, but might include:

- Open standards for query definition.
- Agreed data sets.
- Shared configuration management specification for (bioinformatics) compute.
- Disclosure risk assessment (for ‘filtering’ PHI or identifiable data from results).

The Global Alliance has pioneered a number of the technical prerequisites, including the data sets and query mechanisms, to the extent that a meaningful test can now be constructed to raise confidence in the approach.

Are there downslides to federated analysis?

There is some concern that federated models are too rigid or expensive to implement and don't fulfil the basic criteria of (a) encouraging data owners to participate, and (b) being useful for analyst end users. In commercial settings, such as flight or hotel APIs in travel booking, extensive federated searches exist, so there may be lessons to learn from the success factors in those settings.

The G4AGH proof of concept

The Global Alliance for Genomics and Health GA4GH has been established to promote the required sharing of human genetic data across multiple sites. The Alliance brings together almost 400 expert organisations from across the world in a bid to address some of the sharing and analysis issues which the genomics community faces.

The GA4GH has now put in motion the groundwork for a limited three month proof of concept (PoC), encompassing up to five sites (see Candidate Organisations), that would trial the concept of federated genomic analysis in a bid to address the computational challenges facing the healthcare industry.

Following some discussion at GA4GH Leiden plenary meeting in June 2015. Tests are expected to run in September 2015, with reports expected in October.

Objectives

The primary objectives of the PoC are:

- to increase pragmatic understanding of federation strategies through a set of simulated test cases.
- to document the experience for global community benefit.
- to provide feedback on how the Genomics API model could be extended to capture the requirements of clinical transactions (could be done by way of a second test using SMART-on-FHIR for real-world clinical data formats and an existing API).

Outputs

The PoC will demonstrate and document the experience of a number of sites implementing a first level of API. It is the intention that this improve communication and understanding of how federated analyses might work at a pragmatic level, as well as providing useful feedback to the APIs themselves.

The expected outputs include:

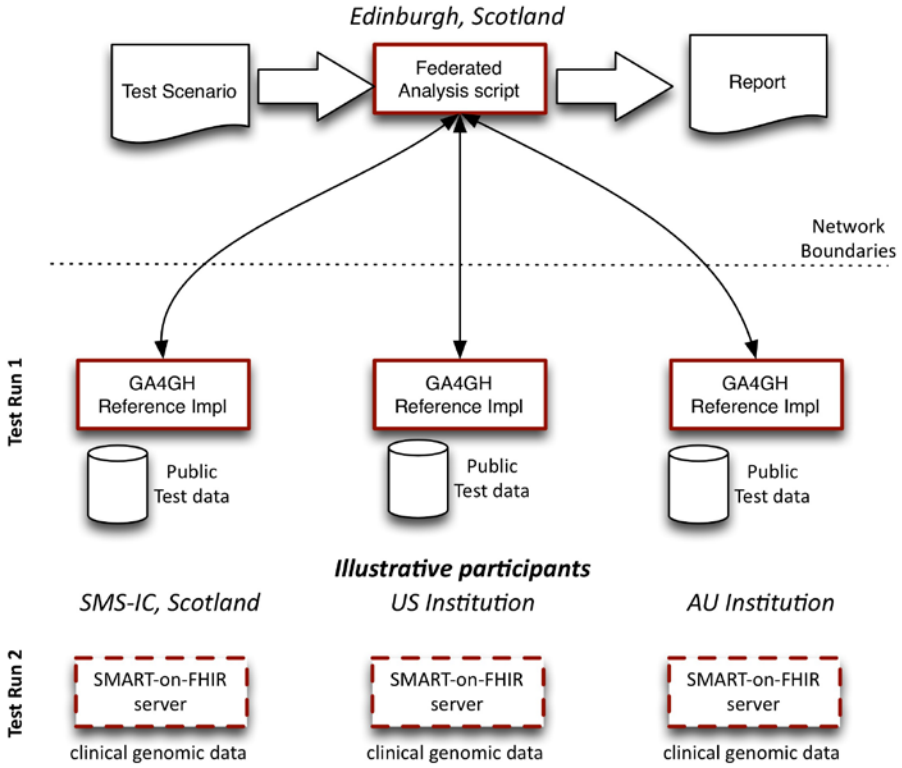
- Installation instructions for conducting a test run of the API.
- A joint report to the Data and Clinical (eHealth) working groups by Aridhia (and Wellcome Trust Centre for Human Genetics, Oxford).
- A presentation at an appropriate GA4GH plenary meeting.
- Feedback and change requests to multiple GA4GH groups including the DWG Genomics API, security working group and CWG eHealth data standards.

Approach

In order to ensure success during the recommended three month duration of the project, the PoC will keep to a tight remit that can be delivered using existing technologies, data and other resources.

The PoC will recruit up to five sites that will deploy reference implementations of the Genomics API (or equivalent) and make them available for a short period of time. The sites will commit to exposing some appropriate public data via the API for a controlled test.

The coordinating group for the PoC will act as an approved researcher or analyst exercising the API across all the sites, as illustrated below.



A set of tasks will be defined for a well understood use case that goes through the different levels of data discovery, aggregation and analysis that will realistically reflect what we would expect a user to require.

We envisage an initial test run with a plan for a second test run based on feasibility:

- Test Run 1: based on the core GA4GH API reference implementation (or equivalent site-specific implementation) and public test data distributed across the sites.
- Test Run 2: based on the GA4GH API and a SMART-on-FHIR Server and some equivalent clinical test data distributed across the sites.

Note that in different settings, the terms 'clinical data', 'clinical genomic data' and 'phenotype' are somewhat overlapping. We use the term 'clinical genomic' to refer to data used for clinical purposes in a genomic workflow. This is likely to include routine EHR data and specific phenotype data collected for that workflow.

Target audience & related programmes

Organisations that operate as 'hubs' for genomic activity (clinical and research), such as the Stratified Medicine Scotland Innovation Centre, will be taking note of this activity and trying to understand how they can participate. Competition and the need to understand information governance issues will encourage participation (see Candidate Organisations).

Some organisations, including those participating in this proposed exercise, have already developed useful models of federated analysis. For example, BioGrid Australia offers an interesting and successful version of federation, where data is shared easily and quickly by authorised researchers accessing data linked permanently in real-time for a project specific purpose.

At the same time, in a clinical context, a number of efforts exist to capture what clinical data should accompany genomic samples for analysis, and what information should be returned after secondary analysis, or after interpretation. The GA4GH Clinical Working Group CWG is leading efforts to share experience and establish standards in eHealth.

For the CWG, and the parallel Data Working Group, this exercise should provide useful feedback from users that can improve the structure of the API and reference implementation, as well as the utility of the data formats.

Technical Support

In order to mitigate any of the barriers to entry that may exist in terms of technical capabilities, the GA4GH has developed a number of technical interfaces (APIs - application programming interfaces) that organisations could implement in order to enable them to participate, through the Data Working Group DWG:

- The GA4GH Genomics API
- The Beacon API
- The MatchMaker project
- Definitions for genomic data types, the Schema.

The Genomics API is defined as:

> *The API is implemented as a webservice to create a data source which may be integrated into visualization software, web-based genomics portals or processed as part of genomic analysis pipelines. It overcomes the barriers of incompatible infrastructure between organizations and institutions to enable DNA data providers and consumers to better share genomic data and work together on a global scale, advancing genome research and clinical application.*

There may be some overlap with the work from the MatchMaker group to define protocols. The PoC working group sees this work as complementary, since it focuses on evolving the technical and scientific protocols based on the Genomics API, whereas the proposed PoC looks at issues of implementation and clinical utility.

GA4GH & related federation approaches - layers

The current portfolio of approaches in the GA4GH can be applied and supplemented to provide layers of analysis in the PoC:

Protocol Level	Purpose
Beacon	Is there data?
Sizing/sampling	Is there enough data?
Interoperability	Is the right kind of data?
Quality check	Does the hub have good enough data?
Privacy preserving aggregates	Run analysis locally, deliver risk-filtered reports, no PHI
Privacy preserving informatics	Provide defined platform as a service to execute complex analysis and deliver risk-filtered outputs
Shared Deposition	Complete or partial results are deposited at participating sites

Technical task definitions

The test case is defined as a typical workflow with a number of 'tasks' (not project tasks, but workflow tasks).

The tasks are defined to reflect a realistic analytical lifecycle and incremental exposure of detail in the data. This should parallel the levels of information governance controls that would be required for organisations to participate and fulfil their existing commitments. Consequently we can ensure that both groups (i.e. both analyst and genomic hub) are engaged in the process.

A scientific goal should be defined which requires the span of API calls in a meaningful ways.

The following table provides an indication of the API calls to be made (most to be confirmed), where the analysis is executed (local, remote) and what the privacy concern level might be for the analysis itself.

Definitions:

- Privacy levels for discussion.
- For detailed reference to APIs - see the API documentation.
- Local - analysis happens on the client machine.
- Remote - a call is made to the API server.

Task group	Location	API	Privacy level?
Discovery	Remote	searchVariants or Beacon?	Medium
Data acquisition - genomic variants	Remote	TBC	High
Data acquisition - phenotype	Remote	TBC	High
Aggregation of data from N sources	Local	TBC	Medium
Exploration	Local	TBC	Low
Analysis (e.g. modelling association)	Local	TBC	Low
Reporting	Local	TBC	High
Citation and acknowledgement	Remote?	Experiment data has some?	Medium?

Candidate Organisations

The following organisations have indicated their participation in the PoC:

Organisation	Location, URL	Points of contact	Role
Stratified Medicine Scotland Innovation Centre	Glasgow/Edinburgh, Scotland http://stratmed.co.uk	Rodrigo Barnes (for Mark Beggs)	API Test site
UCSC	Santa Cruz, California http://www.genome.ucsc.edu	Mark Diekhans	API Test site
Royal Melbourne Hospital & Biogrid Australia	Melbourne, Victoria https://www.biogrid.org.au/	Ingrid Winship	API Test site
Beijing Institute of Genetics, Chinese Academy of Science	Beijing, China http://www.big.ac.cn	Zeng Changqing	API Test site
EMC R&D	Skolkovo, Russia http://www.emc.com	Kamil Isaev, Leonid Levkovich-Maslyuk	API Test site
Wellcome Trust Centre for Human Genetics	Oxford, England http://www.well.ox.ac.uk/home	Jerome Kelleher	Support/Observer
Harvard/MIT	Cambridge, Massachusetts http://hst.mit.edu/	Gil Alterovitz	Support/Observer
Aridhia Informatics	Edinburgh/Glasgow, Scotland www.aridhia.com	Rodrigo Barnes	Coordination Researcher 'proxy'

Participating organisation requirements

Each participating organisation will commit to deploying the API and opening up an internet-facing network address for the specified period of time and at their own cost.

The PoC group will provide instructions on installation of the reference implementation of the API ('the API server') and assist as resources permit to integrate the API server.

The sites are responsible for obtaining permissions to allow access to the data from authorised users for the purposes of the test. Pragmatically, it might be better to use public data that can be freely shared, but this may not be possible or useful. The default plan will be to distribute a reasonable public data set (to be confirmed) among the sites.

Commitment and next steps

1. Each centre should review this proposal and confirm their ability to participate in the exercise pending detailed implementation discussions.
2. It is expected the PoC will require test sites to set aside storage, computing and networking resources to host a reference implementation server and a share of public test data.
3. A telecom will be organised late July to provide more details and agree the plan.
4. There is currently no funding for this exercise, but we hope to make a business case for follow up funding for future test runs and developments.

Resources/Links

- Global Alliance For Genomics And Health GA4GH
- Data Working Group DWG
- Clinical Working Group CWG
- GA4GH Genomics API
- Beacon API
- MatchMaker group

Acknowledgements: Thanks to Jerome Kelleher (Wellcome Trust Centre for Human Genetics, Oxford University) and Professor Andrew Morris (University of Edinburgh), for comments on earlier drafts.