Common Infrastructure for National Cohorts in Europe, Canada, and Africa
- CINECA -

Deliverable D8.1
Report of CINECA Kick-off meeting

Work Package:	WP8 - Project Management and Coordination
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leader:	Thomas Keane
Contributing Partner(s):	EMC
Contractual Delivery Date:	28th February, 2019
Actual Delivery Date:	06 March, 2019
Authors of this Deliverable:	Thomas Keane, Leslie Glass
Dissemination Level:	Public
Type of Deliverable:	Report
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	48 months

Table of contents:

1.Executive Summary

The CINECA consortium was formed in response to the EU call 'Better Health and care, economic growth and sustainable health systems' (H2020-SC1-BHC-2018-2020) with a proposal for an international flagship collaboration with Canada for human data storage, integration and sharing to enable personalised medicine approaches. CINECA proposes a federated cloud enabled infrastructure making population scale genomic and biomolecular data accessible across international borders, accelerating research, and improving the health of individuals across continents. CINECA will leverage international investment in human cohort studies from Europe, Canada, and Africa to deliver a paradigm shift of federated research and clinical applications. The CINECA consortium will create one of the largest cross-continental implementations of human genetic and phenotypic data federation and interoperability with a focus on common (complex) disease, one of the world's most significant health burdens. The partners represent a unique combination of scientific excellence with experience of eleven diverse cohorts and scientific projects such as the European Genome-phenome Archive, CanDIG, and H3Africa.

The CINECA Kick off meeting was held on January 24th-25th 2019 at the Wellcome Genome Campus Conference Centre, Hinxton UK. The key objective of the meeting was to bring together consortium members to facilitate discussion on the project's goals and action plan. The report focuses on an overview of the Work Packages as presented to the consortium (focusing on deliverables due in the first reporting period), the cohorts included in the project, and the decisions made by the Executive Board for actions to implement year 1 of the project.

2. Project objectives

This completion of this deliverable contributes to the following objectives:

    a) All members of the consortium were able to meet face-to-face facilitating discussion on common goals and strategy.

    b) The membership of the consortium was finalised and the budget re-allocated due to the withdrawal of one beneficiary from the original nineteen.

    c) The addition of WP9, an additional ethics work package, was finalised.

3. Detailed report on the deliverable

3.1 Work Package Overviews

3.1.1 WP1: Federated Data Discovery & Queries (WPL: Jonathan Dursi, SickKids)

Jonathan Dursi (SickKids) presented for WP1. He gave an overview of the participants in WP1, and the need to work closely with other WPs, notably WP2 (Interoperable AAI) and WP4 (Federated analyses) regarding operations, as well as a requirement to interact regularly with WP3 (Metadata), 6 (Outreach), 7 & 9 (Ethics). The main goals were outlined as the provision of APIs and portals which use APIs for programmatic/interactive discovery of relevant datasets, and handling complex search queries across cohort datasets.

The first deliverable, a programmatically queryable service catalogue listing all services, is due in Month 18. It is envisaged that for this reporting period the progress will be largely independent of the other WPs. Three risks were identified for which mitigations were also proposed: potential difficulty in reaching agreement on API standards/definitions, problems with the interoperability of implementations, and the emergence of new data types/services not yet supported by APIs. Three key areas of stakeholder engagement were identified/proposed, in that the primary audience for WP1 deliverables in the first reporting period will be other genomics projects interested in interoperable query APIs, WP1 will engage other projects through existent GA4GH workstreams, and they will be planning a hackathon to coincide with upcoming GA4GH meetings.

3.1.2 WP2: Interoperable Authentication and Authorisation Infrastructure (WPL:

Mikael Linden, CSC; Michal Procházka, MU)

Mikael Linden (CSC) presented for WP2. He presented the aim of this WP being to leverage existing ELIXIR and CanDIG AAIs for 3 purposes:

> 1. the creation of the largest interoperability network across Canada, Europe, and Africa 2. to upgrade and enhance the AAIs to support new concepts such as GA4GH researcher

> identifiers to simplify and intensify data sharing 3. to enable at least 10 new cohort infrastructures across Canada, Europe, and Africa to access

federated AAI. The expected impact from the completion of this WP would be that researchers would only require a single set of credentials to access cohorts in all 3 regions, that Data Access Committees would have electronic tools to manage data access applications, and that standardised protocols would be established for mediating access rights to cohorts. The first deliverable will be due at M18, a public demonstration on the integration of European and Canadian AAI systems. The workstream plans for this first reporting period was presented. Risks identified included delays in GA4GH Researcher ID specification, difficulties in understanding individual cohorts' needs or conflicts between those needs,

and delays in AAI implementation. Mitigations were proposed including participation in the GA4GH DURI workstream goals and prioritisation, improving communication, simplification and focus on defining the needs of the respective cohorts and prioritisation where delays are detected. Stakeholder engagement will be initiated with a questionnaire regarding current authentication and authorisation requirements. Initial sustainability is built into this WP as the new features will become part of their hosting infrastructures (e.g. ELIXIR AAI), with standardised protocols such that components can be replaced or operated locally.

3.1.3 WP3: Cohort Level Meta Data Representation (WPL: Helen Parkinson, EMBL-EBI; Fiona Brinkman, SFU)

Fiona Brinkman (SFU) presented for WP3. The two main goals of WP3 are to address metadata representation needs for the cohorts (as aggregate and individual data) across studies and over time, and to develop tools and best practises for the reduction of curatorial burden by creating automated/semi-automated processes for the review and extraction of cohort metadata. The development of an ontology (objective 3.5) is key to this, structured information with a standardised well-defined hierarchy of terms which are interconnected with logical relationships. WP3 will impact on WP1 enabling data discovery by phenotype, WP4 federated joint cohort analyses, and WP5 clinical applications. WP3 will lead to the acceleration of research by reducing curatorial burden for metadata, and it will generate richer, more structured metadata organisation thus enabling insights not previously possible.

The first deliverable due in M18 is the Cohort minimal metadata model. Key factors identified for the completion of this deliverable are learning more about the cohorts and their metadata, which requires access to the cohort metadata in order to initiate. Detailed plans for accomplishing the second deliverable on Semantic Harmonisation (M24) were also presented. Risks identified and mitigations discussed identified early cohort interaction as key with suitable access, learning more about data types, cohort experiences, input from the cohorts on the metadata model, and the ontologies used. A coordinated plan for text mining was identified as requisite. The initial plans for sustainability reflected on EBI's long track record of strong support for metadata harmonisation and development.

3.1.4 WP4: Federated Joint Cohort Analysis (WPL: Kees Van Bochove, THE HYVE BV)

Ward Weistra (THE HYVE BV) presented for the WP4 team. Three major objectives were identified:

1. Development of a toolset for federated analysis of genetic/genomic datasets, while

respecting consent and ELSI framework 2. Development of a toolset for data sources to map to WP3 metadata standards, 3. Development of portable analysis pipelines for finding disease associations based on federated cohort data. The impact of this WP will include discovery of eQTLs, reproducible and proven infrastructure for the 'Personal Health Train' whereby analysis progresses through different datasets without the need to centralise, and personalised medicine based on transnationally-trained algorithms.

The first deliverable due D4.1 (M12) is a report on the trust model for partner sites, and between sites and controlled-access researchers. D4.1 has dependencies on WP3 (machine readable consent ontology) and WP7 (legal ethical framework). Presentation of the timeline for the first reporting period also included detail for Tasks 4.2 and 4.3, Execution APIs and environments for Federated Analyses and Federated Genomic Analyses Algorithm Development. Dependencies identified for 4.2 include internal (WP1, WP2, and WP3) as well as external dependencies of the GA4GH Work Stream and the ELIXIR Compute Platform Environment. Key risks and mitigations identified include:

1. Failure to establish effective technical interfaces between cohort infrastructures and use cases, which could be mitigated by alignment with internal (cross-WP) and external partners via architecture and development calls, virtual hackathons and demonstrations. 2. Datasets not made available for federated research and clinical applications, mitigated by Development synthetic federated data and existing datasets available through European Genome-phenome Archive. 3. Federated technical solutions do not scale to the predicted of 60-500M human genomes, mitigated by testing on synthetic federated data. Stakeholder engagement plans include plans for a Requirements and Architecture Workshop. Initial sustainability plans are that tools would be open source with support by community of service providers, and the alignment with other initiatives such as ELIXIR and GA4GH.

3.1.5 WP5: Healthcare Interoperability and Clinical Applications (WPL: Andrew Stubbs, EMC; Patrick Ruch, HES-SO)

WP5 was presented by Andrew Stubbs (EMC). The four main goals are to develop and evaluate novel health-related applications, deliver a FAIR Federated biomarker discovery and Biobanking Service, to enable clinical research scientist access to big data and supported compute infrastructures, and to improve care planning in healthcare and ensure GDPR compliance. WP5 will impact personalised medicine and translational research by enabling research and diagnostic scientists to access federated services whilst complying with FAIR data principles, and developing utility for Machine- and Deep-learning for disease pathology and drug toxicity prediction.

WP5 deliverables are not due in the first reporting period, and there is an acknowledged dependency on deliverables from WP1-WP4 in order to deliver. Timeline presentation for the first reporting period included work plan for Task 5.2 'FAIR data analysis for cancer biomarker discovery' (M12-M48 led by EMC), and task 5.3 'leveraging federated cohort data to support clinical decision and patient diagnosis' with an example of the scoring service to assess the pathogenicity scales of variants (SIB, M12-M30). Identified risks include the dependencies mentioned on prior WP deliver, that clinical research and diagnostics users have radically different views on the analysis and reporting of results, which will be mitigated by engaging clinical scientists early in the process to deliver user-identified methodologies. An additional risk conceptually is a new data type coming into usage (eg, single cell NGS) that lacks standards and is unsupported by WPs 1-4. Stakeholder engagement will include consultation with the

target groups within clinical research and diagnostic teams internally and externally to EMC and UMCG, seeking input for WP5 services including user specifications, prototype development, and user validation. Stakeholder engagement external to CINECA partners envisaged includes presentation of utility at international meetings and publications, and existing ELIXIR FAIR data analysis collaboration. Sustainability plans include the integration of D5.2 into the ELIXIR Research Service, extending myFAIR to EU members, Canada and worldwide healthcare providers and clinical researchers, and D5.3 integration into RD-connect and the Solve-RD and European Joint Projects for Rare Diseases pushing for national diagnostics consortia.

3.1.6 WP6: Outreach, Training and Dissemination (Cath Brooksbank, EMBL-EBI; Saskia Hiltemann, EMC)

Cath Brooksbank (EMBL-EBI) presented for WP6. The main goals are to raise awareness of CINECA and the opportunities and challenges of sharing cohort data on a global scale, building relationships with related projects, to disseminate the results produced by CINECA through social media channels, blogs, international meetings and publications, and to identify and address training needs by identifying existing learning opportunities and creating new ones both within the consortium itself and more broadly. By involving the technical work packages in outreach, dissemination and training we bring project partners closer to their external stakeholders. The project design ensures this aim because each technical work package has a milestone that feeds into WP6, ensuring that inputs are clear and trackable. External stakeholders will be engaged via surveys, workshops, web, social media, conferences, and consortium-wide participation.

WP6 deliverables due in the first reporting period include D6.1, an Outreach and Dissemination Plan (M09), as well as M12 an annual training programme report (D6.2). D6.1 has identified dependencies requiring the technical WPs to help distribute survey and participate in initial stakeholder meeting, Reykjavik, 23-24 Apr 2019. Delivery of D6.2 in M12 requires technical WPs to deliver on their outreach, training and dissemination milestones. Risks identified include 1-initial stakeholder meeting/survey is poorly attended/poor response rate and not representative of CINECA's stakeholders. These risks can be mitigated by issuing both open and targeted invitations to meeting, including representatives from other key projects/infrastructures, and the timely circulation of programme. Make full use of wider community in GA4GH, H3ABioNet, CANDiG, ELIXIR, BBMRI to circulate survey. Risk 2- Poor uptake of new learning interventions can be mitigated by promoting staff visits extensively at the beginning of the project; advertise short courses etc. via existing networks. Risk 3 - Delays to technical deliverables delay final training deliverables which could be mitigated by setting dates for final course early and in collaboration with relevant WPs; consider online alternatives. An extensive stakeholder engagement plan was outlined including Stakeholder engagement meeting at Global Genomic Medicine Collaborative Meeting, Reykjavik, Iceland, 23/24 April 2019 (Nicky Mulder leading) with participants to be identified by choosing key participants from the CINECA consortium, reaching out to GA4GH, H3ABioNet, CANDiG, ELIXIR, BBMRI and advertising the workshop on the conference website. 2 conference exhibitions planned per year, ideally 1 in Europe, 1 outside Europe, 6 courses, hackathons or knowledge-exchange workshops, based on outcomes of stakeholder analysis, 24 webinars, and 2 staff visits per technical WP.

Sustainability is built into WP6 as the final course on federated analysis of cohort data will be incorporate into EMBL-EBI training programme and development of an online version. Ensure that training is incorporated into GA4GH activities through partner engagement.

3.1.7 WP7 Ethical and legal governance framework for transnational data-sharing (WPL: Michaela Theresia Mayrhofer, BBMRI-ERIC; Emmanuelle Rial-Sebbag, INSERM)

WP7 was presented by Michaela Th. Mayrhofer (BBMRI-ERIC). The main goals of WP7 are to address the practical and strategic ELSI challenges of international research, and to provide the project with practical, accurate and well-grounded ethical and legal recommendations for implementation. These WPs will have a positive impact by contributing to improved collaboration, and by promoting and facilitating the reuse of samples and data for scientific knowledge and gain. Impact is envisaged on a wider stakeholder community including external researchers and policy makers, which will be enabled via enhanced visibility through inclusive expert workshops, and validated policies.

WP 7 deliverables due in the first reporting period a D7.1 Catalogue of ELSI issues V1.0, (M12) and D7.4 DMP – Data Management Plan, (M6) Open Research Data Pilot (BBMRI). A moderate risk identified for WP7 is that GDPR and Canadian/African equivalent laws impede goals of the project. This could impact on all WPs. Therefore a key aim of WP7 is to map the soft and hard law instruments and international policies across Europe, Canada, and Africa to inform the project about potential barriers or impediments to personal data sharing. Stakeholder engagement will be via planned expert workshops, as well as connecting at conferences. Sustainability will be via established routes, eg, BBMRI-ERIC CS ELSI, GA4GH.

3.1.8 WP8 Project Management and coordination (WPL: Thomas Keane, EMBL-EBI)

WP8 was presented by Friederike Bernsdorff (EMBL-EBI). The Project Management team aim is to support the consortium, the Executive Board and the coordinator in the execution of the Grant Agreement and will ensure a smooth running of the project and timely delivery of all project deliverables and milestones. By providing project management expertise, tools and templates to the consortium we will support the smooth operation and successful delivery of a sustainable CINECA project. A key part of this is the production of the Project Handbook, which will be accessible to all members of the consortium via our shared Google Drive folders.

WP8 Deliverables due in the first reporting period include the Kick Off Meeting report D8.1, (M2) and D8.2 the CINECA Risk Registry (M3). Risks in CINECA will be assessed in the areas of data security, ethics, dissemination, innovation, and scientific quality. Identified risks will be recorded in the CINECA risk register and updated regularly with WP leads contributing. Both deliverables have dependencies on all other WPs. Within WP8 there are tasks such as Development and Implementation of a benefit-sharing plan for low and middle income countries/South Africa, (Contributors: UCT), M01-06 which although not

associated with a discrete Deliverable is critical to the success of the project. WP8 identified risks include failure to deliver deliverables and milestones on time. Mitigations include quarterly internal reporting for every WP to facilitate the tracking of progress. Brexit may hamper travel/recruitment. Mitigation is to provide updates as and when available to the consortium.

3.1.9 WP9 Ethical and Legal Compliance (WPL: Thomas Keane, EMBL-EBI; Michaela Theresia Mayrhofer, BBMRI-ERIC; Emmanuelle Rial-Sebbag, INSERM)

WP9 was presented by Emmanuelle Rial-Sebbag (INSERM). The main goals of WP9 are to address the ethical requirements requested by the ethics review, specifically:
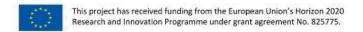
- Humans: the ethical procedures in regards to human participants were not fully detailed to cover the interviews and surveys ; still missing further details regarding to consent procedures.
- Protection of personal data: some additional concerns remain uncovered in regards to structured interviews and surveys that will be done with representative stakeholders

● Third countries: Further details on ethical procedures and material imported/exported are requested. Also there are concerns about Big data export/import material not fully presented in the proposal. The internal impact of this WP is to enable researchers to do science while respecting the EU and the nationals laws (e.g. GDPR) but going beyond the duties to respect the laws we must also identify the questions posed by the development of innovative practices and to build common responses. We aim for standardisation of the processes and elaboration of a common culture for ELSI among the partners. The external impact of this package is to enforce the EU ELSI requirements throughout. Compliance leads to good practices and to the trust of participants.

WP9 deliverables 9.1-10 are all due in the first reporting period. The timeline for the first reporting period was mapped out with three key points identified: at M6 - Humans (consent, ethics approval, previously collected data), M12 - Protection of personal data, M18 - Ethics report developed by Ethics Management Group and reviewed and counter-signed by Independent European Ethics Expert and submitted to the EC at reporting period (M18). However, it was emphasised that the Ethical requirements are not only at the start of the project but during the whole life course of the program where ethical dilemmas may arise and/or new legal instruments may be adopted. Risks identified include dependencies on prompt provision of requested information from partners in order to complete and deliver these deliverables on time. As part of stakeholder engagement to facilitate this it was proposed that each partner nominate an ELSI contact for data collection.

3.2 Cohorts

Critical issues which were identified and discussed included:

- What is the source of the participants?
- What data and phenotypes are available?

- ● What are the plans for the cohort over next 4 years?
- ● What is the model for access+analysis? Can the data be transferred?
- ● Is it possible to identify a single cohort for "end-to-end" tracking for all WPs? An overview of the available cohorts, databases and repositories including location, disease or longitudinal type, and available data types proposed are identified in the following table:

Cohort/Resource name	Number of participants	Location	Longitudinal	Diseases	Gender	WGS	WES	RNA-Seq	Epigenetics	Genotyping
CHILD	3.5K	CA	X	Population based developmental health and disease	M & F	X		X	X	X
CARTaGENE	43K	CA	X	Population based cohort	M & F	X		X		X
PROFYLE	450	CA		Pediatric Oncology	M & F	X		X		
CLSA	50K	CA	X	Population based cohort	M & F					X
H3Africa	75K	SA		Multiple communicable and non-communicable diseases in multiple African countries	M & F	X	X			X
BIOS	4K	NL		Population based cohort	M & F	X		X	X	X
Estonian Biobank	51K	EE	X	Population based cohort	M & F	X	X	X	X	X
CoLaus	6.1K	CH	X	Cardiovascular diseases	M & F			X		X
PsyCoLaus	3.6K	CH	X	Mental disorders	M & F			X		X
EGA	700K	UK+ES		Multiple disease and healthy cohorts	M & F	X	X	X	X	X
UK Biobank	500K	UK	X	Population cohort and disease: cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia	M & F	X	X			X

3.3 Executive Board Meeting

The Executive Board Meeting met on the 25th January and was attended by all members of the EB, or designated representatives for two members with conflicts. The agenda from the EB meeting is included as Appendix 1.

Critical decisions made on the day included finalisation of the membership of the EB, and discussion on membership of the ISAB including new proposals for members. It was proposed that at least one AGM be held in Canada, and the feasibility of holding on in SA to be investigated. The incorporation of Canadian reporting was queried. The importance of the Project Handbook as a resource was discussed. The necessity for timely provision of Risk Register updates was discussed. The implications of the additional ethics requirements was discussed in detail, including the resolution of funding this work. Membership of the Ethics Management Group was resolved. Management of the various communication channels was discussed and responsibilities assigned. Task and budget reallocation resulting from the withdrawal of the Swiss Biobanking Platform from the consortium was agreed. Action items were identified and circulated. Resolution of the AIs are in progress. Complete minutes of the meeting   may   be   found   here: https://docs.google.com/document/d/1veYGDdk8jkzgpzFU_i0X_044fvtogoGOzzfQyWobI9A/edit#hea ding=h.5w50q9hzolq6

4. Abbreviations

AAI -   Authentication and Authorisation Infrastructure
API -   Application Programming Interface
DAC -   Data Access Committee
DURI -   Data Use and Researcher Identities
EB -   Executive Board
ELSI -   Ethical, Legal and Social Issues
eQTL -   Expression quantitative trait loci
FAIR -   Findability, accessibility, interoperability, and reusability
ISAB -   Independent Scientific Advisory Board
GA4GH -   Global Alliance for Genomics and Health
GDPR -   General Data Protection Regulation
NGS -   Next Generation Sequencing

5. Delivery and schedule

The delivery has been delayed by a few days due to holiday conflicts.

6. Appendices

Appendix 1
Agenda

Appendix 2

Attendance List for Kick Off Meeting

Kaur Alasoo	UTARTU
Friederike Bernsdorff	EMBL-EBI
Fiona Brinkman	SFU
Catherine Brooksbank	EMBL-EBI
Tony Burdett	EMBL-EBI

Angel Carreño Torres	CRG
Elisa Cirillo	HYVE
Melanie Courtot	EMBL-EBI
Jonathan Dursi	SickKids
Agnieszka Egan	EMBL-EBI
Éloïse Gennet	INSERM
Leslie Glass	EMBL-EBI
Melanie Goisauf	BBMRI-ERIC
Frédéric Haziza	CRG
Saskia Hiltemann	EMC
William Hsiao	SFU
Thomas Keane	EMBL-EBI
Giselle Kerry	EMBL-EBI
Ilkka Lappalainen	CSC
Mikael Linden	CSC
Mark McCarthy	UOXF
Michaela Th. Mayrhofer	BBMRI-ERIC
Michelle Mendonca	EMBL-EBI
Alex Michie	ClinicaGeno
Nicola Mulder	UCT
Arcadi Navarro	CRG
Tommi Nyrönen	CSC
Helen Parkinson	EMBL-EBI
Michal Prochazka	MU
Jordi Rambla	CRG
William Rayner	UOXF
Emmanuelle Rial-Sebbag	INSERM
Patrick Ruch	HES-SO

14

Serena Scollen   ELIXIR

Dylan Spalding   EMBL-EBI

Andrew Stubbs   EMC

Morris A Swertz   UMCG

Douglas Teodoro   SIB

Ward Weistra   HYVE

Harm-Jan Westra   UMCG