hCNV community led implementation study from 2021 to 2023

Title: Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

Intersecting ELIXIR Platforms, Communities and Projects:

- ELIXIR Galaxy Community
- ELIXIR AAI Infrastructure Service
- ELIXIR Compute Platform
- ELIXIR Training Platform
- ELIXIR FHD Community
- ELIXIR Health Data Focus Group
- ELIXIR Beacon Strategic Implementation Study
- ELIXIR Interoperability Platform

External Projects and Partners:

- EJP-RD
- GA4GH (Discovery, Genomic Knowledge Standards, Phenopackets)

Purpose

The initial 2019-2021 hCNV community implementation study employed a set of perceived needs to a) deliver first community standards and procedures; b) identify intersections with other ELIXIR communities and stakeholders in ELIXIR connected organizations, such as GA4GH; and c) to streamline priorities for relevant, achievable deliveries of hCNV community projects.

This proposal for an hCNV implementation study focuses on those potential high-value targets for data access and delivery, using reference resources and community stakeholder engagement to directly implement and test hCNV resources aligned with ELIXIR ecosystems.

The main target here will be the empowerment of the Beacon protocol, to act as standard for federated hCNV discovery and data delivery, in conjunction with additional GA4GH derived standards.

Node	Name of PI	PMs
ELIXIR-CH	Michael Baudis (MB), Alexander Kanitz (AK)	8
ELIXIR-FR	Christophe Béroud (CB), David Salgado (DS)	3
EMBL-EBI	Timothee Cezard (TC), Kirill Tsukanov (KT)	2
ELIXIRL-NL	Bauke Ylstra (BY)	
ELIXIR-ES	Jordi Rambla (JR), Sergi Beltran (SB), Steven Laurie (SL)	2
ELIXIR-UK	Tim Beck (TB), Anthony Brookes (AB), Krzysztof Poterlowicz (KP)	4
ELIXIR-DE	Björn Grüning (BG)	1
	Total	20
Delivery	Starting from June 2021 for a period of 24 months.	

Abstract

Genomic copy number variations (CNV) are a type of structural genome alterations which represent a major contributor to human genetic variability, genetic disease burden and, importantly, somatic genome variations in cancer. In contrast to precisely defined genome variations (SNV, INDEL and similar), CNV and related variants lack standardized formats for data representation and exchange, a need which was a major driver behind the establishment of ELIXIR's "hCNV Community" in 2018.

Since its creation, the ELIXIR hCNV Community has established itself as a network of active participants from various ELIXIR Nodes and has produced deliverables for the original implementation study, most notably in the areas of file and data exchange formats. Important aspects have been in the close interaction with the ELIXIR Beacon project and with other ELIXIR Communities such as Galaxy, allowing us to prioritize the work of the hCNV Community as well as the identification of relevant partner initiatives in ELIXIR Communities and beyond (e.g. in GA4GH Work Streams).

The main objective of this proposal is to reinforce work on priority areas established in the current hCNV Implementation Study, and extend collaborations with the Rare Diseases and Galaxy Communities, as well as both EJP-RD and GA4GH. Outcomes will consist of :

- shared CNV resources testing advanced versions of the Beacon protocol, thereby demonstrating its utilization as well as driving protocol development
- the integration of GA4GH standards such as Phenopackets in such resources
- tools for data ingestion and export for standard formats (e.g. VCF, Phenopackets)

and CNV-specific improvements of such standards

- demonstration of ELIXIR AAI implementations on clinical and research hCNV resource instances
- demonstration of Galaxy pipeline adoption for real-world hCNV data analysis projects
- dissemination through training to trigger the adoption of the tools

This Implementation Study is composed of 5 Work Packages:

1. Establishment of a community of genomic data resource providers from different technical and content scopes (e.g. cancer, rare diseases, raw data repositories) for iterative assessment and removal of data access limitations
2. Implementation of current or forward-looking versions of the Beacon protocol as baseline standard on participating resources, for practical evaluation of protocol limitations, with close feedback into Beacon development
3. Interaction with the ELIXIR Galaxy Community to enable data flow between hCNV reference resources and Galaxy workflows
4. Assessment of hCNV related workflows and tools for integration with data exchange and processing procedures, for hCNV resources and community platforms
5. Support of hCNV related training and documentation throughout ELIXIR Communities, integrating with the ELIXIR Training Platform

Description of Work

WP1 - hCNV community reference resources

Lead	ELIXIR-CH (MB)
Members	ELIXIR-FR, EMBL-EBI, ELIXIR-UK, ELIXIR-ES
Delivery	**M1-M24**

The interest of "hCNV" as data format is spread over a number of partially diverse types of biomedical data resources:

- diagnostics and resources in clinical genetics, focussed on genotype/phenotype associations, predominantly in the "rare diseases" area - ELIXIR-FR & -UK
- large-scale data submission repositories - EMBL-EBI
- research-centric data aggregation resources with specific focus on structural variations e.g. in cancer - ELIXIR-CH

All these different scenarios are represented throughout the participating ELIXIR members as labeled above.

The focus of the work package will be to assemble a set of actively participating hCNV resources and stakeholders, representing these different general scopes, to evaluate domain-specific options and limitations for general hCNV (and related) data access, and to provide a set of hCNV resources of different scopes to the ELIXIR communities.

Deliverables:

D1.1 Deliverable (**M1-12**): Document with description of current status, roadblocks and envisioned solutions for all participating hCNV resources
D1.2 Deliverable (**M13-24**): hCNV resource hub as entry point for the different resource types (connected to WP2)

Milestones:

M1.1 Milestone (**M12**): Documentation of status (D1.1) for re-evaluation of project plan

WP2 - hCNV Resources and Beacon

Lead	ELIXIR-CH (MB, AK), ELIXIR FR (DS)
Members	ELIXIR-ES, ELIXIR-UK (TB, AB)
Delivery	**M1-M24**

The Beacon protocol - developed with leadership from ELIXIR since 2016 - is emerging as the de facto development target for federated discovery of genomic variants. While the Beacon protocol has supported CNV queries from its originally approved version (GA4GH Beacon v1), the version 2 of the protocol currently under development will add specific CNV and range query types as well as options for biomedical and technical metadata queries.

The objective of this work package is to perform implementation-based testing of Beacon query scenarios on hCNV resources, using emerging protocol options, and to interact with the Beacon team for perceived additions to the protocol where necessary. Items we have identified as potential candidates include e.g. queries for named genome elements (genes, repeat regions etc.) and their implementation in the Beacon protocol or possibly through service layers, as well as fine-grained CNV queries and responses (e.g. ploidy levels, data provenance, annotations).

Deliverables:

D2.1 Deliverable (**M1-12**): Demonstration of at least one public Beacon implementation with extended hCNV functionality and adoption of ELIXIR infrastructure concepts (e.g. AAI)

D2.2 Deliverable (**M12-24**): Integration of multiple (i.e. more than one, from different providers) hCNV Beacons in ELIXIR Beacon network, with functionality beyond current (v1.n) positional requests and rich data delivery (e.g. Phenopackets)

Milestones:

M2.1 Milestone (**M12**): Documentation of Beacon implementation and hCNV Beacon roadmap

WP3 - Galaxy Community Intersection and Data Exchange

Lead	ELIXIR-UK (KP), ELIXIR-DE (BG)
Members	ELIXIR-CH (MB), ELIXIR-FR (DS)
Delivery	**M1-M24**

Following some preliminary discussions, an increasing interest of the Galaxy community with respect to use and delivery of hCNV reference data as well as the integration of hCNV related data handling has been identified. This work package will specifically address the potential use of data from participating hCNV stakeholders' resources in Galaxy community projects and workflows, e.g. as reference datasets in genomic profiling studies. Another aspect will be in the bi-directional exchange of expertise and tools for hCNV specific processing needs. However, the direct development of tooling or benchmarking for Galaxy pipelines is seen as external to this specific implementation study itself.

Deliverables:

D3.1 Deliverable (**M1-6**): Overview of hCNV specific tools and their representation - or lack of - in Galaxy workflows

D3.2 Deliverable (**M7-18**): Establishment and execution of hCNV-Galaxy tool & data exchange procedures

D3.3 Deliverable (**M19-24**): Documentation and revised definition of hCNV-Galaxy

Community interaction goals

Milestones:

M3.1 Milestone (**M12**): hCNV data from at least one of the participating resources and representing a substantial amount of its content, for use in Galaxy Community resources

M3.2 Milestone (**M24**): Report about achieved hCNV tool and data representation inside Galaxy platforms

WP4 - Workflows and Tools for hCNV Data Exchange Procedures

Lead	EMBL-EBI (TC, KT)
Members	ELIXIR-CH (MB, AK)
Delivery	**M1-M24**

Due to the lack of practical, standardized annotation and file formats for hCNV and related structural genome variants, current hCNV resources and analysis pipelines typically rely on customized data formats for data storage and exchange (e.g. custom BED-like columnar formats). Some file standards like VCF—the reference file format for exchange of called genome variants—in principle allow for the annotation of CNVs. However, interpretation of VCF file content for structural variants has been shown to be fraught with ambiguities (e.g. precise INDEL vs. CNV; relative copy number interpretation, interpretation of multi-allelic variants and such with polyploid baseline).

This work package will aim at improving representation of SVs and CNVs in file formats for community data exchange (such as VCF), as well as providing documentation, example files and validation tools for hCNV optimised versions of genomic file formats.

Deliverables:

D4.1 Deliverable (**M1-12**): Status report of current handling, limitations, needs of CNV specific file formats.

D4.2 Deliverable (**M13-24**): Updated status report of current handling, limitations, needs of CNV specific file formats.

Milestones:

M4.1 Milestone (**M24**): Report of implementation of improvements to community data exchange file formats and validation tools.

WP5 - Training and dissemination

Lead	ELIXIR-CH (MB)
Members	ELIXIR-FR (CB, DS), ELIXIR-UK (AB, TB)
Delivery	**M1-M24**

The "Training and Dissemination" work package is aimed at two major directions:

- the representation of the ELIXIR hCNV project and community to potential outside interactors such as GA4GH work streams and projects or EJP-RD
- the generation, dissemination and visibility of hCNV related topics and resources throughout ELIXIR, e.g. by co-opting existing training frameworks (TESS)

As part of the outside representation, this work package will include the continuous maintenance of the hCNV online representation (https://hcnv.github.io).

Deliverables:

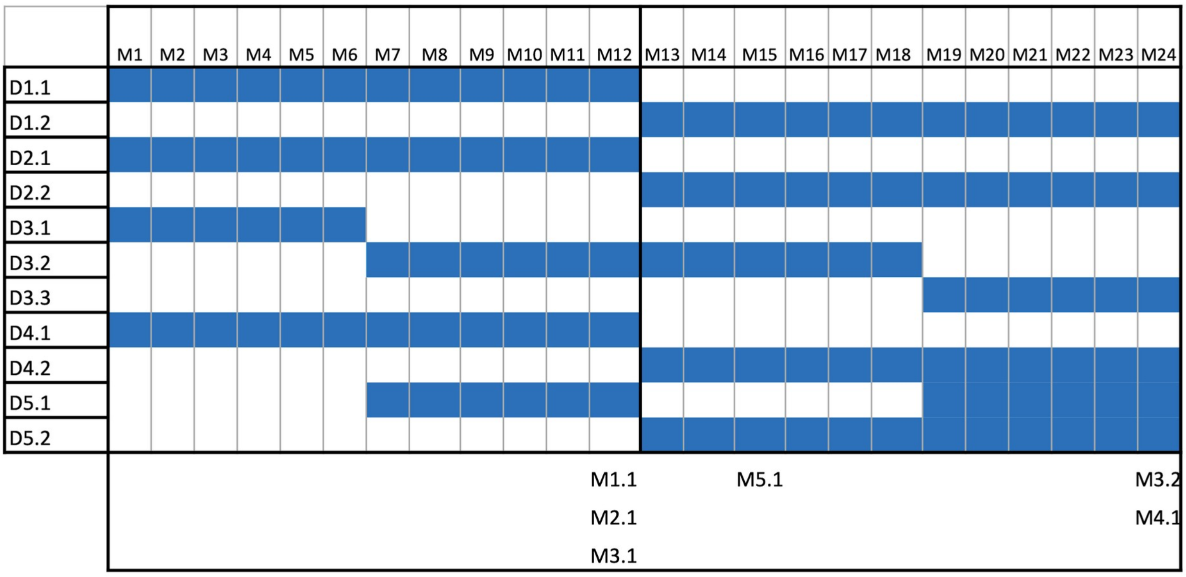
D5.1 Deliverable (**M1-6 & 19-24**): Demonstration of ongoing/repeated representation of the project in ELIXIR and outside events, documented on th ehCNV community website

D5.2 Deliverable (**M13-M24**): Coordinate CNV training efforts between all ELIXIR-hCNV initiatives

Milestones:

M5.1 Milestone (**M15**) hCNV "Code & Data" community event ("Contentathon"-timepoint represents establishment of planning)

Gantt chart of Milestones and Deliverables



	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24
D1.1	■	■	■	■	■	■	■	■	■	■	■	■												
D1.2													■	■	■	■	■	■	■	■	■	■	■	■
D2.1	■	■	■	■	■	■	■	■	■	■	■	■												
D2.2													■	■	■	■	■	■	■	■	■	■	■	■
D3.1	■	■	■	■	■	■																		
D3.2							■	■	■	■	■	■	■	■	■	■	■	■						
D3.3																			■	■	■	■	■	■
D4.1	■	■	■	■	■	■	■	■	■	■	■	■												
D4.2													■	■	■	■	■	■	■	■	■	■	■	■
D5.1							■	■	■	■	■	■							■	■	■	■	■	■
D5.2													■	■	■	■	■	■	■	■	■	■	■	■
												M1.1 M2.1 M3.1			M5.1									M3.2 M4.1

The Gantt chart depicts the principle nature of the project as ongoing community engagement effort with a limited set of milestones.

Budget

https://docs.google.com/spreadsheets/d/1mX8muPzD07aoaJcufJVmvtQKMN23xY2Xf0W-If01Bog/edit#gid=0

Alignment with evaluation criterion 1: Scientific focus, scope, need

Genomic copy number variations (CNV) are a major, poorly understood, contributor to germline variation and disease susceptibility, and ubiquitous elements of neoplastic transformation. A bottleneck in their interpretation is the lack of common standards for CNV annotation and exchange, and therefore CNV data is frequently excluded from comparative analyses and variant interpretation resources.

New data sharing and federation protocols - most notably based on the Beacon API or platforms such as Galaxy - promote the discovery and interpretation of genomic variants across multiple resources, using standardized schemas and methods. However, the practical application of CNV discovery through Beacon so far has been driven by few reference projects (e.g. progenetix.org) and upcoming Beacon versions will need use case driven extensions for efficient query and representation of structural variants. Also, there is a need for hCNV data and tooling in communities such as Galaxy, with high potential to further synergistic developments.

Alignment with evaluation criterion 2: Community served

The strength of this proposal lies in synergies between different communities within ELIXIR, but also beyond. Main benefits - previously identified but in need to become initiated through this project - are between hCNV & Galaxy. Data & tools from the hCNV community will be delivered to expand Galaxy use cases; the Galaxy Community can contribute workflows and efficient resource use for hCNV analysis scenarios. For Beacon development, hCNV community members will be able to design & test "beyond Beacon" implementations, i.e. such built around the Beacon protocol but testing its use-case driven expansion, and feed this back to Beacon development while becoming agile adopters of the protocol.

Overall, community intersections have been identified for Galaxy Community, AAI Infrastructure Service, Compute Platform, Training Platform, FHD Community, Health Data Focus Group, Beacon Strategic Implementation Study and Interoperability Platform.

Importantly, interaction with outside communities and organizations - namely GA4GH - will be a direct result of the project.

Alignment with evaluation criterion 3: Quality of service

The proposed project does not deliver a service itself, but will drastically improve the power of dedicated ELIXIR community and project based services, through alignment, testing and exchange of standards for genomic and related metadata exchange, the distribution of reference and research data between communities and the expanded community interactions as well as the increased awareness associated with such networked activities.

Alignment with evaluation criterion 4: Supporting the mission of ELIXIR

The main components of this project center around open standards development and data exchange, for biomedical research and application use cases. By its nature, the project will improve data accessibility as well as more efficient data re-use and bioinformatics resource utilization.

Importantly, with tight integration to international communities such as GA4GH and its participants through standards and protocol development, this project will contribute to increased visibility and impact of ELIXIR activities and ELIXIR driven standard developments.

Alignment with evaluation criterion 5: Sustainability and impact

The project is aimed chiefly to drive the increased utility of standards and resources, thereby contributing to increased impact and sustainability of ELIXIR resources. It delivers per se no resource itself which has to be maintained or is limited by the extent of the study. However, communities and documentation (e.g. through the community's website) will benefit from follow-up projects and community members are expected to seek continuation support.