ELIXIR: h-CNV community implementation study from 2019 to 2021

Purpose

The human CNV community (h-CNV) has been officially created in December 2018. It aims to address the major challenge of NGS data interpretation in the era of whole genome sequencing for the most frequent mutation type: Copy Number Variation. Seven topics have been identified during the kick-off meeting and further refined with all h-CNV partners. This ultimately led to the proposal described in this implementation study.

Node	Name of PI
ELIXIR-FR	Christophe Béroud, David Salgado, Marc Hanauer, Victoria Dominguez
ELIXIR-CH	Michael Baudis
ELIXIR-DE	Jan Korbel
EMBL-EBI	Thomas Keane, Fiona Cunningham
ELIXIR-ES	Joaquin Dopazo, Alfonso Valencia, Salvador Capella, Sergi Beltran, Steven Laurie, Gemma Bullich, Laura I. Furlong, Janet Piñero
ELIXIR Hub	John Hancock, Gary Saunders, Kathi Lauer, Leyla Garcia
ELIXIR-NL	Bauke Ylstra, Daoud Sie, Leon Mei, Morris Swertz (UMCG), Lennart Johansson
ELIXIR-NO	Eivind Hovig, Pubudu Samarakoon
ELIXIR-HU	Attila Gyenesei ,Katalin Monostory
ELIXIR-SI	Brane Leskošek, Polonca Ferk, Marko Vidak
ELIXIR-UK	Krzysztof Poterlowicz
Delivery	Starting from June 2019 for a period of 24 months.

Description of Work

WP1 - Optimal CNV detection pipelines for research and diagnostics

Lead	Alfonso Valencia, Salvador Capella (**ELIXIR-ES**)
Members	**CH** (Michael Baudis), **DE** (Jan Korbel), **ES** (Joaquin Dopazo, Steven Laurie, Gemma Bullich), **FR** (Christophe Béroud and David Salgado), **HU** (Attila Gyenesei), **NL** (Bauke Ylstra and Daoud Sie, Leon Mei, Lennart Johansson), **SI** (Brane Leskošek, Polonca Ferk, Marko Vidak), **UK** (Krzysztof Poterlowicz),
Delivery	**M1-M18**

Multiple publications have reported pipelines to detect CNV using micro-array, WES or WGS data. Nevertheless, there is a low consensus among the tools in calling CNVs, especially from widely used WES experiments: a moderate sensitivity (50 to 80%); a fair specificity (70 to 94%) and a poor false discovery rate (27 to 60%). This underlines that read-depth based programs are still immature for WES-based CNV detection with a low sensitivity and an uncertain specificity. Comparable experiences were revealed by participants of the ELIXIR h-CNV workshop, and it was concluded that, even if micro-array technologies provide overall better CNV detection parameters, the wide adoption of NGS technologies represents a true challenge for the accurate detection of CNV. Based on these observations, the need for an extensive assessment for research and diagnostics and benchmarking of existing tools, within OpenEbench ELIXIR infrastructure, for somatic and/or germline CNV detection in gene panels, WES, shallow and deep WGS, array CGH and SNP arrays was established as one of the working areas of the h-CNV community. The WP1 objective will be to release a set of sensitive and reliable pipelines, optimized and validated to detect CNV from various high throughput datasets. These pipelines will be available either through the ELIXIR compute nodes and/or as stand-alone solutions.

Milestones:

M1.1 Evaluation of available systems to detect CNV and documentation in ELIXIR Bio.tools (**M6**)

M1.2 Installation of systems to be benchmarked within ELIXIR compute platform and OpenEbench, work on metrics to be collected and compared (**M6**)

M1.3 Proceed to benchmark and provide results (**M12**)

M1.4 Benchmark systems to detect CNV for diagnostic (ISO) requirements. (**M18**)

Deliverables:

D1.1 Deliver the list of available pipelines/software as well as partners' local solutions to detect CNV from gene panels, WES, shallow and deep WGS, array CGH and SNP arrays. (**M6**, **M12**, **M18**)

D1.2 Develop a generic benchmarking platform to evaluate new tools and new datasets. (**M9**, **M18**)

D1.3 Benchmark the various systems using datasets from WP2 to select the most sensitive, specific, reliable and rapid systems for each dataset for germline and somatic CNVs. (**M12**, **M18**)

D1.4 Deliver optimized pipelines from D1.3 to increase performance on ELIXIR compute nodes and define optimal parameters and guidelines to help end-users to efficiently and reliably detect CNV in various situations through the ELIXIR training platform. (**M12**, **M18**)

D1.5 Genomic external quality assessments (EQAs) for CNV patient sample data, germline and somatic. (**M18**)

WP2 - Definition of reference datasets

Lead	Steven Laurie (**ELIXIR-ES**); Cristina Y. Gonzalez (**EMBL-EBI**)
Members	**DE** (Jan Korbel), **EMBL-EBI** (Thomas Keane), **FR** (Christophe Béroud and David Salgado), **SI** (Marko Vidak)
Delivery	**M1-M15**

The ambition is to provide open reference datasets of fully validated somatic and germline CNVs representing a wide range of samples types and experimental technologies. These reference materials will be available to the community to evaluate and compare pipelines and/or NGS technologies and/or for quality assurance. This will include both digital raw data and, potentially, biomaterials.

To do so, the h-CNV community will align with international initiatives such as Genome in a Bottle (GIAB) and the Global Alliance for Genomics and Health Benchmarking Team to establish reference datasets including various CNV (deletions and duplications) of various sizes ranging from a single exon CNV to large genomic rearrangements. Two subsets will be defined for germline and somatic CNVs. These datasets will contain samples with fully validated CNVs by other approaches such as Multiplex ligation-dependent probe amplification (MLPA) and quantitative or semi-quantitative PCR. During the community kick-off meeting, participants discussed the GDPR and its impact on reference human datasets. The participation of lawyers and ethics specialists is therefore needed, and this was proposed to be addressed at a more global level by the ELIXIR Human Data Communities as a whole. In case no human reference dataset could be exchanged at a community level, alternatives using other model organisms has been proposed. As previously mentioned, the NGS technologies are rapidly evolving and therefore the reference datasets will be updated at M15.

Milestones:

M2.1 List of reference datasets for NGS (**M3**)

M2.2 Updated list of reference datasets for NGS (**M15**)

Deliverables:

D2.1 Deliver WES reference datasets (**M6**, **M15**)

D2.2 Deliver WGS reference datasets (**M6**, **M15**)

WP3 - Improvement of community formats for CNV exchange

Lead	Thomas Keane (**EMBL-EBI**)
Members	**CH** (Michael Baudis), **FR** (Marc Hanauer), **EMBL-EBI** (Cristina Y. Gonzalez), **SI** (Brane Leskošek).
Delivery	M1-M12

International collaborative projects require harmonization and standardization of results in order to ensure efficient data aggregation and comparison. Although various international initiatives, such as the GA4GH Genomic Knowledge Standards and Large-Scale Genomics groups, are currently globally addressing aspects of this issue, no robust and exhaustive standard CNV annotation format has emerged so far. The h-CNV partners discussed the adoption of the VCF format and its current limitations for the CNV field. It was concluded that, although this format is well-known by molecular biologists and could therefore be a starting point, it is less frequently used by cytogeneticists. There is therefore a strong need to improve the VCF format and identify other nomenclatures and widely used formats in other communities. It is being recognized that any development or improvement of standards for CNV annotation should be performed in alignment with existing efforts, notably GA4GH work streams and ELIXIR interoperability platform.

Milestones

M3.1 Perform survey to identify community use of file formats and storage formats for representation and exchange of CNV data. (**M6**)

Deliverable:

D3.1 Catalog of identified issues and limitations of file formats and data schemas for representing and exchanging CNV data. (**M6**)

D3.2 Create consensus collection of perceived requirements for efficient and effective CNV file and data exchange formats. (**M6**)

D3.3 Propose and communicate specification changes to pre-existing formats (e.g. VCF) to address the highest priority limitations. (**M9**)

WP4 - Enabling CNV data discovery in diagnostic and phenotypic context

Lead	Michael Baudis (**ELIXIR-CH**), Marc Hanauer (**ELIXIR-FR**)
Members	**DE** (Jan Korbel), **EMBL-EBI** (Denise Carvalho-Silva), **ELIXIR-Hub** (Gary Saunders, Leyla Garcia), **ES** (Alfonso Valencia, Salvador Capella, Joaquin Dopazo, Laura I. Furlong, Janet Piñero, Steven Laurie, Gemma Bullich), **FR** (Christophe Béroud and David Salgado, Marc Hanauer), **HU** (Katalin Monostory), **NL** (Bauke Ylstra), **SI** (Marko Vidak, Polonca Ferk), **UK** (Krzysztof Poterlowicz)
Delivery	**M1-M12**

For many types of genomic data analyses, a main target is the quantitative and qualitative association of individual germline and somatic genome variants to sample associated characteristics, such as clinical diagnosis, population background, geographical and environmental measures and clinical parameters. Finding similar cases at the clinical level is a key component of clinical diagnosis and research, to identify disease-causing genes and to explain genotype/phenotype correlations.

Within the Discovery workstream of GA4GH standards, such as Beacon, are being developed to establish data discovery across federated networks of databases using common application programming interfaces (API) and associated reference implementations. The “ELIXIR Beacon” project - a GA4GH “Driver Project” - allows data owners to add their resources to “Beacon Networks” through straightforward implementation of predefined schemas and tools (https://beacon-project.io). With the recent addition of CNV representation to the Beacon API and planned ontology-based phenotype queries, the Beacon ecosystem represents a prime target for the implementation of (CNV related) genotype-phenotype representation and querying. The h-CNV group will work towards enabling its use for the envisioned patient discovery, through the support of extended clinical descriptions including enabling and testing of relevant annotation standards (e.g. HPO, NCIt and additional ontologies). ELIXIR h-CNV partners are already strongly involved in the development of ontologies such as HPO and ORDO and in the mapping of various ontologies, medical terminologies, vocabularies and nomenclatures, and will contribute their recommendations to standards proposed by GA4GH (clinical and phenotypic data capture) workstream. In addition, national databases described in WP6 will adopt the recommendations from WP4 to ensure cross queries and the identification of similar patients.

Milestones:

M4.1 List of selected ontologies to capture phenotypic description of patients and samples (**M12**)

Deliverables:

D4.1 Deliver a list of select ontologies required to efficiently capture phenotypic description useful for data interpretation for any genetic disease. (**M6**)

D4.2 Deliver lists of common data elements that should be provided in various situations such as rare diseases, oncology or common diseases. (**M6**)

WP5 - Creation of innovative tools

Lead	Christophe Béroud and David Salgado (**ELIXIR-FR**), Joaquin Dopazo (**ELIXIR-ES**)
Members	**CH** (Michael Baudis), **DE** (Jan Korbel), **EMBL-EBI** (Sarah

	Hunt), **ES** (Laura I. Furlong, Alfonso Valencia, Salvador Capella), **HU** (Katalin Monostory), **NO** (Eivind Hovig, Pubudu Samarakoon), **UK** (Krzysztof Poterlowicz)
Delivery	**M1-M24**

CNVs can involve large genomic regions and affect multiple genes. With relevance to recessive diseases and tumor suppressor genes, CNVs can co-operate with other types of genomic or regulatory alterations affecting alleles of the same genetic target. In situations where large CNV are involved, it is difficult to unanimously identify the specific gene(s) whose alterations are directly associated to the patient's phenotype.

Here, the h-CNV community will develop innovative tools, specifically regarding: Functional annotation of CNVs; combinatorial approaches to CNV interpretation; and identification of landmark genes in regions of interest.

Milestones:

M5.1 Creation of innovative tools to facilitate CNV interpretation (**M12**, **M24**)

Deliverables:

D5.1 Deliver mandatory CNV annotations including: type; genotype; genes and transcripts; expression level; exons; regulatory elements; breakpoints/fusion fragments. (**M6**)

D5.2 Creation of a specific pipeline to interpret duplications as tandem, inverted or translocation duplications may result in very different phenotypes. (**M12**)

D5.3 Creation of specific bioinformatics tools to select candidate genes localized in the CNV region by combining genes' annotations and patients' phenotype. (**M18**)

D5.4 Deliver tools to determine (tumor) heterogeneity and mosaicisms. (**M24**)

WP6 - FAIRification of h-CNV databases and datasets

Lead	Christophe Béroud and David Salgado (**ELIXIR-FR**), Joaquin Dopazo (**ELIXIR-ES**)
Members	**CH** (Michael Baudis), **EMBL-EBI** (Cristina Y. Gonzalez), **FR** (Victoria Dominguez Del Angel), **HU** (Katalin Monostory), **NL** (Morris Swertz), **NO** (Eivind Hovig, Pubudu Samarakoon, Lennart Johansson), **UK** (Krzysztof Poterlowicz)
Delivery	**M1-M24**

Various national CNV databases, curated CNV data resources, and ELIXIR deposition databases are currently being developed by ELIXIR h-CNV partners. In order to allow interoperability (including resource and data discovery), the FAIR principles (Findable, Accessible, Interoperable, Reusable) will be applied to those systems to demonstrate the feasibility and utility of distributed CNV databases. This will respect databases' ownerships and national regulations' compliance while allowing searching for similar patients across the network. To respect and follow these principles, we will ensure that data are:

(i) Findable

- The data should contain globally unique, resolvable and persistent identifiers.
- Include machine-readable descriptions to support structured search and filtering.

(ii) Accessible

- Metadata has to be accessible beyond the lifetime of the digital resource.
- Clearly defines regarding the condition for access and security protocols for sharing data.

(iii) Interoperable

- Usage of standardized formats.
- Extensible machine interpretable formats for data and metadata (e.g. YAML files, JSON-LD).
- Use vocabularies (ontologies) and link with other robust resources.
- Integration with FAIR resources.

(iv) Reusable

- Provide licensing, provenance and description on community-standards.

Milestones:

M6.1 Release of a FAIR CNV database. (**M18**)

Deliverables:

D6.1 FAIRification of the French BANCCO database (http://bancco.fr) developed at Aix Marseille University, the CIBERER (Spanish network for research in rare diseases) database developed at the Fundación Progreso y Salud of Sevilla and the VKGL CNV database (Dutch clinical genetics diagnostics) CNV database as prototypes to demonstrate the benefits of using the FAIR data principles for CNV in diagnostic and research contexts. (**M18**)

D6.2 Extension of the FAIRification to other non-specific CNV databases such as the European Variation Archive (EVA), RD-CONNECT, arrayMap, Progenetix and others. (**M24**)

WP7 - Dissemination

Lead	Michael Baudis (**ELIXIR-CH**), Victoria Dominguez Del Angel (**ELIXIR-FR**), Gary Saunders (**ELIXIR-Hub**)
Members	**ELIXIR-Hub** (Kathi Lauer), **ES** (Joaquin Dopazo), **HU** (Attila Gyenesei), **NO** (Eivind Hovig, Pubudu Samarakoon), **SI** (Brane Leskosek), **UK** (Krzysztof Poterlowicz)
Delivery	**M1-M24**

The global adoption of tools and guidelines is strongly linked to the ability to communicate, produce training materials and train actors as well as patients and the general public. In addition, capacity building training events will be organized across ELIXIR nodes being less proficient in h-CNV.

Milestones:

M7.1 Participation to international meetings to promote the h-CNV community (**M12**, **M24**)

M7.2 Organisation and participation to international bio hackathons/Jamborees/Capacity building events dedicated to CNV (**M18**)

Deliverables:

D7.1 Creation of Jamborees to gather experts' point of view on the various objectives and related tasks and developments. **(M12, M24)**

D7.2 Creation of regular hackathons to ensure smooth developments and benchmarks by various ELIXIR Nodes. **(M12, M24)**

D7.3 Promotion of the ELIXIR h-CNV community through participation to international meetings, such as GA4GH. A contact has already been established with the Human Genome Variation Society (HGVS). **(M12, M24)**

D7.4 Creation of regular Capacity building events to ensure knowledge dissemination across ELIXIR Nodes. **(M12, M24)**

D7.5 Set Up documentation about best practices and guidelines on CNV interpretation **(M18-M24)**