hCNV community led implementation study from 2021 to 2023

Title: Reference hCNV datasets, use-case workflows and benchmarking

Purpose

The ELIXIR human CNV Community (hCNV) was created in December 2018. In two years contributions to the field have been numerous (ELIXIR IS, Rare Diseases, Federated Human Data, Beacons, GA4GH, EJP-RD and Beyond 1 Million Genomes - B1MG). The Community now aims to address the major challenge of NGS data interpretation in the era of whole genome sequencing: Copy Number Variation. During the first commissioned service offered as a starting grant, the Community has identified various gaps to proceed with CNV tools benchmarking and in particular for Exome and targeted sequencing, which are by far the most widely used technologies in diagnostic laboratories and in research. Within this implementation study we want to provide solutions and bioinformatic infrastructure solutions to fill identified gaps, and to make these biomedical reference materials available (i.e. via Open Science) to the various communities and platforms.

Node	Name of PI	PMs
ELIXIR-FR	Christophe Béroud (CB), David Salgado (DS)	9
ELIXIR-UK	Krzysztof Poterlowicz (KP)	4
ELIXIR-CH	Michael Baudis (MB)	2
ELIXIR-ES	Salvador Capella-Gutierrez (SCG)	3
ELIXIR-ES	Sergi Beltran (SB), Steve Laurie (SL)	In kind
ELIXIR-DE	Björn Grüning (BG)	2
	Total	20
Delivery	Starting from June 2021 for a period of 24 months.	

Interactions and utility to other projects

ELIXIR platforms:
Data, Tools, Interoperability, Training

ELIXIR Communities:
hCNV, Galaxy, Rare diseases, Federated Human Data

National and International projects:
EJP-RD, B1MG, EOSC-Life, EOSC-Pillar

Abstract

During the first Community led Implementation study, the ELIXIR hCNV community has identified that only limited datasets exist to test and benchmark tools for the analysis of CNV and structural variations. Furthermore, recent datasets focused on high-quality Whole Genome Sequencing (WGS) analyses but not on the most commonly used Whole Exome Sequencing (WES) or genomic array technologies. The **generation of publicly accessible reference sets** (raw and interpreted CNV data) for a variety of technological platforms will allow the hCNV community to generate the mandatory material. It will allow the **creation of "control datasets"** required by many detection tools, and **to complement standardization and benchmarking efforts** such as the "Genome in a Bottle" initiative.
This implementation Study is composed of 4 Work Packages: "Dataset selection and generation"; "Analyse and Compare CNV with other benchmarking initiatives"; "Exploitation of the datasets by the Galaxy Community"; "Use-case driven Services Bundles for ELIXIR communities and platforms"; and "Training and dissemination". They will cover all aspects from samples' selection and data generation to their use by the ELIXIR Data, Tools, Interoperability, and Training platforms and ELIXIR communities as the Galaxy, and the ELIXIR Human Data Communities.

Description of Work

WP1 - Dataset selection and generation

Lead	ELIXIR-FR (CB, DS)

Members	ELIXIR-ES (SCG)
ELIXIR platforms to interact with	ELIXIR Data Platform
Delivery	**M1-M18**

Copy Number Variations correspond to the presence of a specific genomic region in various numbers in a population. The most frequent alleles correspond to duplication or deletions but multi-copies can also be observed. These alleles can either have a phenotypic impact or be silent. Their detection has for a long time been difficult due to technologies' limitations and their positioning between molecular biology and cytogenetics. Nowadays, they can be routinely detected by microarrays and sequencing. Both approaches have limits in their specificity and sensitivity. To improve data analysis pipelines, it is required to access reference materials with well-characterized CNV of various sizes and types.

The goal of this Work Package is to add to the existing reference materials for WGS, reference materials for WES and gene panels. To do so, we will first select the relevant individuals from the Personal Genome project, which have already been analyzed by microarrays, short and long reads WGS technologies ("Genome in a Bottle" - GIAB). Moreover, these datasets can be key material for national or international initiatives, for instance based on recent discussion, this material can serve as demonstrator for the B1MG WP4 - Federated platform

As the most widely used WES CNV analysis pipelines compare samples to reference individuals (≥10), it is mandatory that the new datasets contain data from at least 15 individuals. In addition, as the WGS technologies directly sequence genomic DNA without prior capture of specific regions, the coverage is more uniform and therefore cannot be used to generate artificial WES / gene panels datasets.

Taking into account the requirements for research and diagnostic laboratories accreditation procedures, it is required that the samples can be acquired from a cell-line provider.

In collaboration with the ELIXIR data platform and ELIXIR partners involved in EOSC-Life and EOSC-pillar projects, we will identify the best strategy to release and provide access to these datasets.

Deliverables:

D1.1 Deliverable (**M3**): Identify the 15 biological samples for which cell-lines (and potentially WGS/microarrays data) are available (ELIXIR-FR).

D1.2 Deliverable (**M8**): Generation of WES reference datasets from the 15 individuals by two sequencing kits (ELIXIR-FR).

Milestones:

M1.1 Milestone (**M18**): WES/gene panels reference datasets are available to the general public through the ELIXIR data platform and deposition database (ELIXIR-FR).

WP2 - Analyse and Compare CNV with other Benchmarking initiatives

Lead	ELIXIR-ES (SCG), ELIXIR-FR (CB, DS)
Members	ELIXIR-UK
ELIXIR platforms to interact with	ELIXIR Tools platform
Delivery	**M1-M18**

The main objective of this WP is to make use of the datasets identified and/or generated in WP1 to generate reference datasets specifically designed for benchmarking purposes. To be able to generate and provide such reference datasets, a set of widely used CNV analysis pipelines will be used to generate an initial set of calls. Different strategies will be applied to generate a high quality set of calls, which will be compared with those from other efforts e.g. Zook et al. Nature Biotech, 2020.

This initial effort will allow us to specifically characterize datasets for different benchmarking efforts. This is particularly relevant for calling CNVs using WES and gene-panels. Thus, datasets will be made available to the community through long-term archive platforms like Zenodo and/or EUDAT B2Drop and deployed into platforms like OpenEBench to promote the benchmarks of new analytical pipelines in connection with WP3. Moreover, these datasets can be subject to further validations

and/or curation to increase their quality and, therefore, their usefulness for the whole CNV community.

Deliverables:

Deliverable D2.1 (**M12**): Initial set of CNVs identified for each sample (ELIXIR-FR).

Deliverable D2.2 (M15). Initial release of high-quality datasets for benchmarking activities through appropriate repositories (ELIXIR-ES).

Deliverable D2.3 (M18). Definition of the relevant metrics to measure the scientific performance of analytical pipelines using data generated by WES and gene-panels with different kits/protocols (ELIXIR-ES)

Milestones:

M2.1 Milestone (**M12**): List of detected CNV in each sample of the reference datasets are available to the community (ELIXIR-FR).

WP3 - Exploitation of the datasets by the Galaxy Community

Lead	ELIXIR-UK (KP)
Members	ELIXIR-DE (BG), ELIXIR-FR (CB, DS), ELIXIR-ES (SCG)
ELIXIR platforms to interact with	ELIXIR Tools platform, ELIXIR Interoperability platform
Delivery	**M1-M24**

In order to facilitate testing, reproducibility and reusability of tools, we will take advantages of the Galaxy project and shared workflows. In the context of CNVs, we will take advantage of the newly generated datasets to automatically benchmark containerized individual tools and containers-based workflows. To promote the broad adoption of community developed platforms e.g. WorkflowHub, OpenEBench, Galaxy; the plan is to create BioContainers and Galaxy integrations for individual tools and deposit workflows in the WorkflowHub. Galaxy will be used to analyze a set of reference input datasets in a reproducible and transparent way. Resulting

output will be automatically submitted to OpenEBench for the evaluation against the existing benchmark datasets built in WP2.

In this work package we would like to concentrate on the definition of new Structural Variant (SV) and CNV calling tools, and gather benchmark results in an automated fashion.

Deliverables:

Deliverable D3.1 (M12) Provide a number of structural genomic variant calling tools in Galaxy through biocontainers registry process (ELIXIR-UK)

Deliverable D3.2 (M18) Deployment of a fully automated and continuous benchmarking mechanism within OpenEBench to evaluate existing, e.g. updated, and newly developed tools and workflows for specific CNV analyses (BSC, ELIXIR-ES).

Milestones:

Milestone M3.1 (M8) Define the process of biocontainers registry, tools integration into Galaxy, and then benchmarking for CNV tool (ELIXIR-UK)

Milestone M3.2 (**M14**) Guidelines on how to provide datasets to OpenEBench for the automated benchmarking of containerized tools and workflows deposited in WorkflowHub and executed using Galaxy (BSC, ELIXIR-ES).

WP4 - Training and dissemination

Lead	ELIXIR-UK (KP)
Members	ELIXIR-FR (CB, DS), ELIXIR-ES (SCG), ELIXIR-CH (MB)
ELIXIR platforms to interact with	ELIXIR Training platform
Delivery	**M1-M24**

In this work package, we will produce training materials, notes and documentations to illustrate key steps of this Implementation study.

Our objective will be through engagement with the Galaxy community and a range of virtual Contentathons to build Galaxy Training Network (GTN) resources on CNV data analysis using Galaxy tools and workflows delivered in WP3 and referenced datasets delivered in WP1 of the study We will disseminate advances and achievements with other ELIXIR communities through participation to ELIXIR events and meetings, and communicate these efforts with outside stakeholder communities and organisations (e.g. GA4GH).

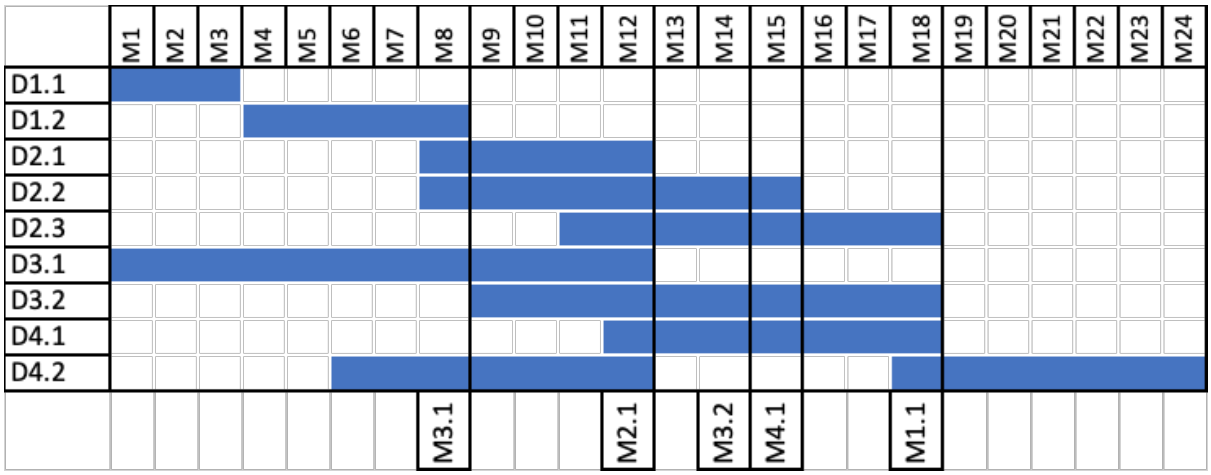
Deliverables:

D4.1 Deliverable (**M18**): Deliver Galaxy training network tutorial (slides, hands on, workflows) on CNV data analysis (ELIXIR-UK)

D4.2 Deliverable (**M12-M24**): Coordinate CNV training efforts between all ELIXIR-hCNV initiatives (ELIXIR-CH)

Milestones:

M4.1 Milestone (**M15**): Organisation of virtual Contentathons to build GTN content (ELIXIR-UK)



Gantt Chart for the Reference hCNV datasets, use-case workflows and benchmarking

Budget :

https://docs.google.com/spreadsheets/d/1jjLaHMwLEFCmrcbNspMf3q79woL9aqLlNZEsBBZlb0c/edit?usp=sharing

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

The hCNV Community aims to address major challenges of NGS data interpretation in the era of whole genome sequencing for Copy Number Variation. During the first commissioned service offered as a starting grant, the Community has identified various gaps to proceed with CNV tools benchmarking and in particular for Exome and targeted sequencing, which are by far the most widely used technologies in diagnostic laboratories and in research. Within this implementation study we want to provide solutions and bioinformatic infrastructure solutions to fill these gaps, and to make these biomedical reference materials available (i.e. via Open Science) to the various communities and platforms. Key deliverables of this IS will be: generation of biological sample sequencing datasets; workflows; their integration within galaxy and study the feasibility of interconnected services between Galaxy workflows for CNV benchmarking and OpenEbench).

Alignment with evaluation criterion 2: Community served

The deliverables of this implementation study will be serving many ELIXIR communities and platform services. The release of biological samples sequencing data will not only open the way to new projects within the hCNV community but will provide key materials for the Galaxy community that will be able to give access to real datasets for training purposes. The datasets and all associated analysis will serve the Rare Diseases and other human data communities by providing guidance on how to analyze CNV data and other type of mutations (SNP, indels).
Finally we believe, based on recent participation to meetings from the ELIXIR Federated human data community, EJP-RD and the B1MG (in particular WP4) projects, that such material will greatly facilitate the generation of demonstrators of genomic data sharing initiatives, without requiring the generation of synthetic datasets that do not fully represent real data.

Alignment with evaluation criterion 3: Quality of service

The proposed project does not deliver a service *per se*, but will produce reference datasets and know-how to various ELIXIR communities and future projects.

Alignment with evaluation criterion 4: Supporting the mission of ELIXIR

All advances made in this IS will be developed in the context of FAIR and Open Science. We will generate datasets, workflows, tools containers, benchmarking results and guidelines that will be reusable for any community and project within and outside the ELIXIR ecosystem and environment. For these purposes we will work in close collaboration with the ELIXIR Hub representatives, ELIXIR platforms and communities. We believe that results from this IS will benefit to research and diagnostic communities in their day to day practice with NGS data.
Additionally, this project is a collaborative project that will tighten relationship between two ELIXIR communities and multiple ELIXIR platforms

Alignment with evaluation criterion 5: Sustainability and impact

We will make sure to provide access to data and resources generated during this project in long term sustainability repositories. We will work with ELIXIR Data and Tools platforms to identify the most relevant Datawarehouse, repositories for containers and workflows catalogues. The produced datasets will eventually be updated overtime depending on technologies evolutions.

This project is unique as it proposes a full journey in the data life cycle. It encompasses data generation, short- and long-term storage, usability and reusability of the generated data and open access. It will provide unique reference resources for CNV data analysis for WES and targeted sequencing experiments. This will be complementary to the WGS reference datasets made available by the National Institute of Standards and Technology from the US and therefore will place EU at the forefront of CNV research.