ELIXIR hCNV

First Implementation Study and Ongoing Work

Michael Baudis | ELIXIR Human Data Communities | 2022-03-15
Why hCNV Community?

Structural Genome Variation Data :: Resources and Technologies

- structural genome variations are a major contributor to genetic diseases and cancer
- knowledge about and standards for copy number variations / aberrations (CNV/CNA) has not been in step with NGS & GWAS driven SNV/SNP assessment

Mission statement

Despite the fact that Copy Number Variations are the most prevalent genetic mutation type, identifying and interpreting them is still a major challenge. The ELIXIR human Copy Number Variation (hCNV) Community aims to implement processes to make the detection, annotation and interpretation of these variations easier.

CNV with unknown clinical impact in a case of Silver-Russel Syndrome

Local Affymetrix Genotyping 6 signal distribution pattern and segmentation result in patient SR12 (SR0012, orange data) and his father (SR0012V, steelblue data). In both samples a duplication in the DSCR can be observed, affecting the whole KCNE1 and DSCR1/RCAN coding regions. In contrast, DYRK1A lays ~2.5 Mb distal of the duplication. Only the genes discussed in this article are shown.
Why hCNV Community?

Structural Genome Variation Data :: Resources and Technologies

- structural genome variations are a major contributor to genetic diseases and cancer
- knowledge about and standards for copy number variations / aberrations (CNV/CNA) has not been in step with NGS & GWAS driven SNV/SNP assessment

Mission statement

Despite the fact that Copy Number Variations are the most prevalent genetic mutation type, identifying and interpreting them is still a major challenge. The ELIXIR human Copy Number Variation (hCNV) Community aims to implement processes to make the detection, annotation and interpretation of these variations easier.
ELIXIR hCNV Community

Structural Genome Variation Data :: Resources and Technologies

- First meeting of group in 2018
- ELIXIR Human Copy Number Variation (hCNV) approved in 2019
- Initial implementation study (2019-2021) for community set-up, gap analysis and exploration of technical deliverable

Mission statement
Despite the fact that Copy Number Variations are the most prevalent genetic mutation type, identifying and interpreting them is still a major challenge. The ELIXIR human Copy Number Variation (hCNV) Community aims to implement processes to make the detection, annotation and interpretation of these variations easier.

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.

<table>
<thead>
<tr>
<th>Node</th>
<th>Name of PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELIXIR-FR</td>
<td>Christophe Béroud, David Salgado, Marc Hanauer, Victoria Dominguez</td>
</tr>
<tr>
<td>ELIXIR-CH</td>
<td>Michael Baudis</td>
</tr>
<tr>
<td>ELIXIR-DE</td>
<td>Jan Korbel</td>
</tr>
<tr>
<td>EMBL-EBI</td>
<td>Thomas Keane, Fiona Cunningham</td>
</tr>
<tr>
<td>ELIXIR-ES</td>
<td>Joaquin Dopazo, Alfonso Valencia, Salvador Capella, Sergi Beltran, Steven Laurie, Gemma Bullich, Laura I. Furlong, Janet Piñero</td>
</tr>
<tr>
<td>ELIXIR Hub</td>
<td>John Hancock, Gary Saunders, Kathi Lauer, Leyla Garcia</td>
</tr>
<tr>
<td>ELIXIR-NL</td>
<td>Bauke Ylstra, Daoud Sie, Leon Mei, Morris Swertz (UMCG), Lennart Johansson</td>
</tr>
<tr>
<td>ELIXIR-HU</td>
<td>Attila Gyenesei ,Katalin Monostory</td>
</tr>
<tr>
<td>ELIXIR-SI</td>
<td>Brane Leskolek, Polonca Ferk, Marko Vidak</td>
</tr>
<tr>
<td>ELIXIR-UK</td>
<td>Krzysztof Poterłowicz</td>
</tr>
</tbody>
</table>

Delivery
Starting from June 2019 for a period of 24 months.
hCNV Implementation Study 2019-2021
Setting the Scope | Solidifying the Community | First Deliveries

• challenge participants and define the wider landscape as well as future directions

• set of 7 work packages
  ➡ landscape analysis
  ➡ technical products
  ➡ resource improvement
  ➡ community building & outreach

• regular meetings, website, hackathons...

  ➤ WP1 - Optimal CNV detection pipelines for research and diagnostics
  ➤ WP2 - Definition of reference datasets
  ➤ WP3 - Improvement of community formats for CNV exchange
  ➤ WP4 - Enabling CNV data discovery in diagnostic and phenotypic context
  ➤ WP5 - Creation of innovative tools
  ➤ WP6 - FAIRification of h-CNV databases and datasets
  ➤ WP7 - Dissemination
hCNV Implementation Study 2019-2021
Setting the Scope | Solidifying the Community | First Deliveries

- highly ambitious goals, beyond available support
  - especially reference / benchmarking dataset generation and pipeline development
- emerging interactions and collaborations with ELIXIR platforms & communities and beyond
  - Galaxy
  - GA4GH / ELIXIR Beacon project

- WP1 - Optimal CNV detection pipelines for research and diagnostics
- WP2 - Definition of reference datasets
- WP3 - Improvement of community formats for CNV exchange
- WP4 - Enabling CNV data discovery in diagnostic and phenotypic context
- WP5 - Creation of innovative tools
- WP6 - FAIRification of h-CNV databases and datasets
- WP7 - Dissemination
hCNV Implementation Study 2019-2021

Some Achievements and Deliveries

- Benchmarking tools and OpenEbench TransBioNet testing event
- Demonstration of CNV detection tools in clinical (cancer) setting
- Amending bio.tools for extensive list of CNV related analysis tools
  - https://bio.tools/t?domain=elixir-hcnv
- Updating / registering shared hCNV resources at fairsharing.org
- Consensus collection of perceived requirements for efficient and effective CNV file and data exchange formats
With the ELIXIR Tools Platform: Bio.tools & EDAM ontology

Within the first commissioned service granted to the community (First hCNV Community IS)
The community created a list of 245 CNV detection tools for various detection technologies NGS (WGS, WES, panel), CGHarray, ...

- We wanted to share this list of tools → Bio.tools
- Started to collaborate with ELIXIR tools platform members (Jon Ison / Hervé Menager)
- We created a specific bio.tools subdomain and listed/annotated (with EDAM terms) 109/245 CNV tools http://elixir-hcnv.bio.tools/
- We contributed to the EDAM ontology to include about 20 specific terms to describe CNV and Structural variations in (topics/operation branches)
Some Achievements and Deliveries

- HGVS satellite meeting – Human CNV – June 14th 2019 – Göteborg Sweden

- hCNV community workshop ELIXIR All-Hands Lisbon – June 2019

- survey of data annotation formats, including comments on VCF development

- start FAIRification of CNV national / reference databases (BANCCO, Progenetix)

- Community white paper published

- Biohackathon Paris 2019

- in 2021 start of shared meetings of subgroup with Beacon variants scout team
Some Achievements and Deliveries

- survey about genomic variation file formats and their use, suitability for representing CNV data
- part of the survey was focused specifically on VCF, a key GA4GH standard at the intersection of human and computer readable formats
- Results
  - BED-like formats are frequently used, but the better defined flavours are not optimal for CNVs and other SVs
  - JSON w/ schema has potential, but still misses finalized GA4GH schemas (VRS emerging) and suffers "readability" issues for non-bioinformatics customers
  - VCF was considered as a/the variant standard file format, but not "CNV-friendly" in v4.2 and in the existing tools for I/O handling of CNV data
hCNV Implementation Study 2019-2021

Some Achievements and Deliveries

- Close interaction with Beacon "scout" teams
  - use case driven (BANCCO - RD & Progenetix - cancer) development of essential query standards for the upcoming Beacon v2
Beacon v2: Extended Variant Queries

Range and Bracket queries enable positional wildcards and fuzziness

- Genome Range Queries provide a way to “fish” for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, reference or alternate bases) limit the scope of the responses
- New Beacon v2 size parameters to limit structural variants (e.g. “focal” CNVs)

- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- Allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- Typical use case is e.g. the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions
**Links with other projects**

- **Galaxy**
- **Beacon**
- **Human data communities**

**Platforms**

**Tools**
- Data *
- Interoperability *
- Training *

* Links to be strengthened through future IS

**National links reinforcements**
- BANCCO national CNV db for diagnostic
- Links to national sequencing projects (PFMG, GOLD)
- ...

**Participation to various taskforces**
- Variant representation
- Beacon
- Future of VCF
- Adopting new standards (phenopackets, DUO)

**Access to CNV resources (Beacon)**
Demonstrator for HOOM ontology

**CNV workshop**
organisation, Götheborg Sweden 2019

**LINKS to be reinforced**
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.

Intersections 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

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

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.

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)
hCNV Implementation Studies 2021-2023 No. 1

Reference hCNV datasets, use-case workflows and benchmarking

- only limited datasets exist to test and benchmark tools for the analysis of CNV and structural variations
- 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
- 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
- creation of “control datasets” required by many detection tools
- complement standardization and benchmarking efforts such as the “Genome in a Bottle” initiative
- integrate with Galaxy community & platforms

- WP1 - Dataset selection and generation
- WP2 - Analyse and Compare CNV with other Benchmarking initiatives
- WP3 - Exploitation of the datasets by the Galaxy Community
- WP4 - Training and dissemination
• reinforce work on priority areas established in the current hCNV Implementation Study
• extend collaborations with the Rare Diseases and Galaxy Communities, EJP-RD and GA4GH
• Expected outcomes
  ➡ shared CNV resources testing advanced versions of the Beacon protocol
  ➡ 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
  ➡ ELIXIR AAI demo on clinical and research hCNV resources
  ➡ demonstration of Galaxy pipeline adoption for real-world hCNV data analysis projects
• connecting to international partners, e.g. Cancer Genomics Consortium (U.S.)

➤ WP1 - hCNV community reference resources
➤ WP2 - hCNV Resources and Beacon
➤ WP3 - Galaxy Community Intersection and Data Exchange
➤ WP4 - Workflows and Tools for hCNV Data Exchange Procedures
➤ WP5 - Training and dissemination
Ongoing...

hCNV & Intl. Community

- contributions to ontologies and standard definitions
- close ongoing interactions with GA4GH work streams
- influencing the development of the GA4GH VRS variant standard

hCNV Community
Genomic Copy Number Variations in Humans

News & Events
ELIXIR All Hands 2022 - hCNV
Representation
CNV Ontology Proposal - Now Live at EFO
CNV Site now at cnvar.org
CNV Implementation Study 2021/2: Beacon and Beyond

Participants
Standards and Guidelines
Studies & Resources
Examples, Guides & FAQ
Contacts
Related Sites
h-CNV @ ELIXIR
Beacon Project
Beacon @ ELIXIR
Schemas Blocks
Github Projects

h-CNV

CNV Ontology Proposal - Now Live at EFO

As part of the hCNV-X work, related to "Workflows and Tools for hCNV Data Exchange Procedures" and the intersection with Beacon and GA4GH VRS - we have now a new proposal for the creation of an ontology for the annotation of (relative) CNV events. The CNV representation ontology is targeted for adoption by Sequence Ontology (SO) and then to be used by an updated version of the VRS standard. Please see the discussions linked from the proposal page. However, we have also contributed the CNV proposal to EFO where it has gotten live on January 21.

Everybody is welcome to contribute to the editing of the proposal at the SO & VRS Github repositories!

2021-01-21: copy number assessment term tree now live on EFO

The copy number assessment term tree has been accepted into the Experimental Factor Ontology and can be used for referencing CNV types.

More ontologies...
... with h-CNV contributions...

Larrybabbs commented 18 days ago

Draft Relative Copy Number class proposal

--- the target region/feature
subject: region/feature/allele/haplotype

→9 quantifiable values that correspond to the EFO copy number assessment term:

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>copy-neutral</td>
<td>neutral loss of heterozygosity</td>
</tr>
<tr>
<td>1</td>
<td>high-level copy number gain</td>
<td>high-level copy number gain</td>
</tr>
<tr>
<td>2</td>
<td>low-level copy number gain</td>
<td>low-level copy number gain</td>
</tr>
</tbody>
</table>

RelativeCopyNumber

Relative Copy Number Variation captures a classification of copies of a molecule within a system, relative to a baseline. These types of Variation are common outputs from CNV callers, particularly in the somatic domain where Absolute Copy Counts are difficult to estimate and less useful in practice than relative statements.

Computational Definition

The relative copies of a Molecular Variation, Feature, Sequence Expression, or a CURIE reference against an unspecified baseline in a system (e.g. genome, cell, etc.).

Information Model

Some RelativeCopyNumber attributes are inherited from Variation.

Field               | Type         | Limits | Description
---                 |--------------|--------|------------------------
_id                 | CURIE        | 0.1    | Variation id. MUST be unique within document.
_type               | string       | 1.1    | MUST be "RelativeCopyNumber"
_subject            | Molecular Variation | 1.1 | Subject of the Copy Number object
_relative_copy_class | string       | 1.1    | MUST be one of "complete loss", "partial loss", "copy neutral", "low-level gain" or "high-level gain".

• original 2019-2021 implementation study provided visibility and established connections for new studies
• instrumental were Biohackathons, use case & standards surveys and co-participation of group members
• future work plans to leverage the resources of participants through pre-established interactions and synergies
• 2 independent studies provide clearer definitions of deliverables and individual scopes