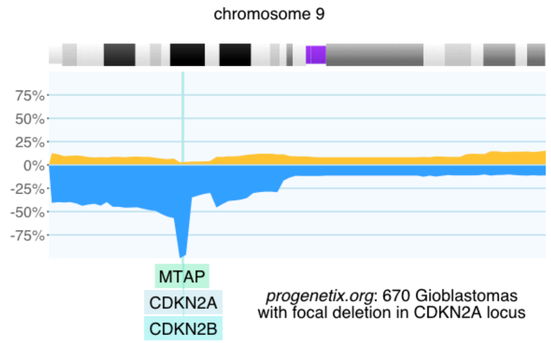
ELIXIR hCNV

Updates and future plans

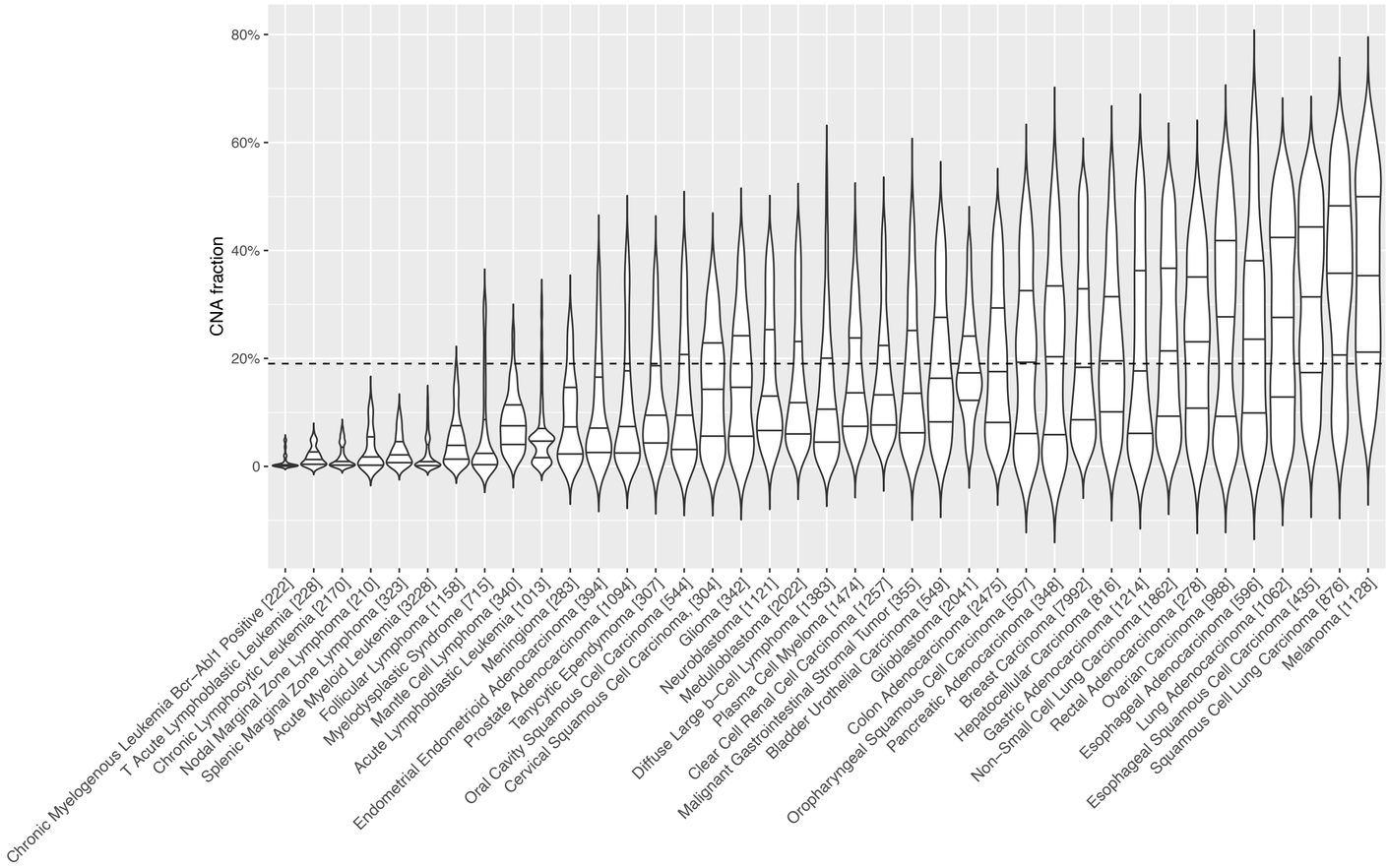
Michael Baudis | ELIXIR Human Data Day | 2021-06-11



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

progenetix

Genome CNV coverage in Cancer Classes

- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)knowledge about and standards for copy number variations / aberrations (CNV/CNA) has not been in step with NGS & GWAS driven SNV/SNP assessment

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.

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.

Christophe Béroud
(ELIXIR France)

David Salgado
(ELIXIR France)

Gary Saunders
(Human Data Coordinator, ELIXIR Hub)

Michael Baudis
(ELIXIR Switzerland)

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 datection 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

hCNV Implementation Study 2019-2021

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

F1000Research F1000Research 2020, 9(ELIXIR):1229 Last updated: 01 JUN 2021

Check for updates

OPINION ARTICLE

The ELIXIR Human Copy Number Variations Community: building bioinformatics infrastructure for research [version 1; peer review: 1 approved]

David Salgado[1], Irina M. Armean[2], Michael Baudis[3], Sergi Beltran[4,5], Salvador Capella-Gutierrez[6,7], Denise Carvalho-Silva[2,8], Victoria Dominguez Del Angel[9], Joaquin Dopazo[10], Laura I. Furlong[11], Bo Gao[3], Leyla Garcia[2,12,13], Dietlind Gerloff[14], Ivo Gut[4,5], Attila Gyenesei[15], Nina Habermann[16], John M. Hancock[13], Marc Hanauer[17], Eivind Hovig[18,19], Lennart F. Johansson[20], Thomas Keane[2], Jan Korbel[16], Katharina B. Lauer[13], Steve Laurie[4], Brane Leskošek[21], David Lloyd[13], Tomas Marques-Bonet[22], Hailiang Mei[23], Katalin Monostory[24], Janet Piñero[11], Krzysztof Poterlowicz[25], Ana Rath[17], Pubudu Samarakoon[26], Ferran Sanz[11], Gary Saunders[13], Daoud Sie[27], Morris A. Swertz[20], Kirill Tsukanov[2], Alfonso Valencia[6,7,28], Marko Vidak[21], Cristina Yenyxe González[2], Bauke Ylstra[29], Christophe Béroud[1,30]

[1]Aix Marseille Univ, INSERM, MMG, Marseille, France
[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[3]Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 4, Barcelona 08028, Spain
[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[6]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[7]Spanish National Bioinformatics Institute (INB)/ELIXIR-ES, Barcelona, Spain
[8]Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
[9]Institut Français de Bioinformatique, UMS3601-CNRS, CNRS, Paris, France
[10]Clinical Bioinformatics Area, Fundación Progreso y Salud, CDCA, Hospital Virgen del Rocio, Sevilla, Spain
[11]Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences, Pompeu Fabra University (UPF), Barcelona, Spain
[12]ZB MED Information Centre for Life Sciences, Cologne, Germany
[13]ELIXIR Hub, Hinxton, UK
[14]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg
[15]Szentágothai Research Center, University of Pécs, Pécs, Hungary
[16]Genome Biology, European Molecular Biological Laboratory, Heidelberg, Germany
[17]Orphanet, INSERM, Paris, France
[18]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[19]Centre for bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway
[20]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[21]Faculty of Medicine - ELIXIR Slovenia, University of Ljubljana, Ljubljana, Slovenia
[22]Institute of Evolutionary Biology (UPF-CSIC), Catalan Institution for Research and Advanced Studies, Barcelona, Spain
[23]Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
[24]Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary
[25]Centre for Skin Sciences, University of Bradford, Bradford, UK

hCNV Implementation Study 2019-2021

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 th eexisting tools for I/O handling of CNV data

ELIXIR hCNV 2019-21 Deliverable D3.2

Project Title:	First hCNV Community Implementation Study
Deliverable title:	Create a consensus collection of perceived requirements for efficient and effective CNV file and data exchange formats
WP No.	3
WP Title	Improvement of community formats for CNV exchange
Contractual delivery date:	30.11.2019
Actual delivery date:	12.12.2019
WP leads:	Thomas Keane
Partner(s) contributing to this deliverable:	EMBL-EBI

Report authors: Kirill Tsukanov[1], Sundararaman Venkataraman, Giselle Kerry, Thomas Keane (EMBL-EBI)

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 Scouts: Genomic Variants Use Cases & Examples

This document develops a set of genomic variant types and associated query formats to be supported by t
Beacon protocol. The initial development focuses on the possibly limited, but unambiguous definition of que
formats, driven and documented through real-world use cases.

DUP (Duplication)

Definitions:

- SO:0001742 - A sequence alteration whereby the copy number of a given region is greater than the reference sequence (copy number gain).
- CNV query resolves to DUP or DEL or CNV (and their equivalents -> see DUP, DEL), tandem dups?
- CNV loss / CNV gain)

Examples below based on a specific study (https://doi.org/10.1016/j.tjog.2018.06.018)

Example: Find duplications involving the whole locus (chr2:54,700,000-63,900,000)

Provided by: David Salgado & Michael Baudis
Notes:

- This is an application of a **Bracket Query**
- Here, matched duplication events start 5` of the region and end 3` of it.
- Besides the positions, this requires knowledge about the maximum value of the reference base (or use of a very large one exceeding chromosome size; this example here uses a lazy "just bigger than chr2" value).

Query structure:

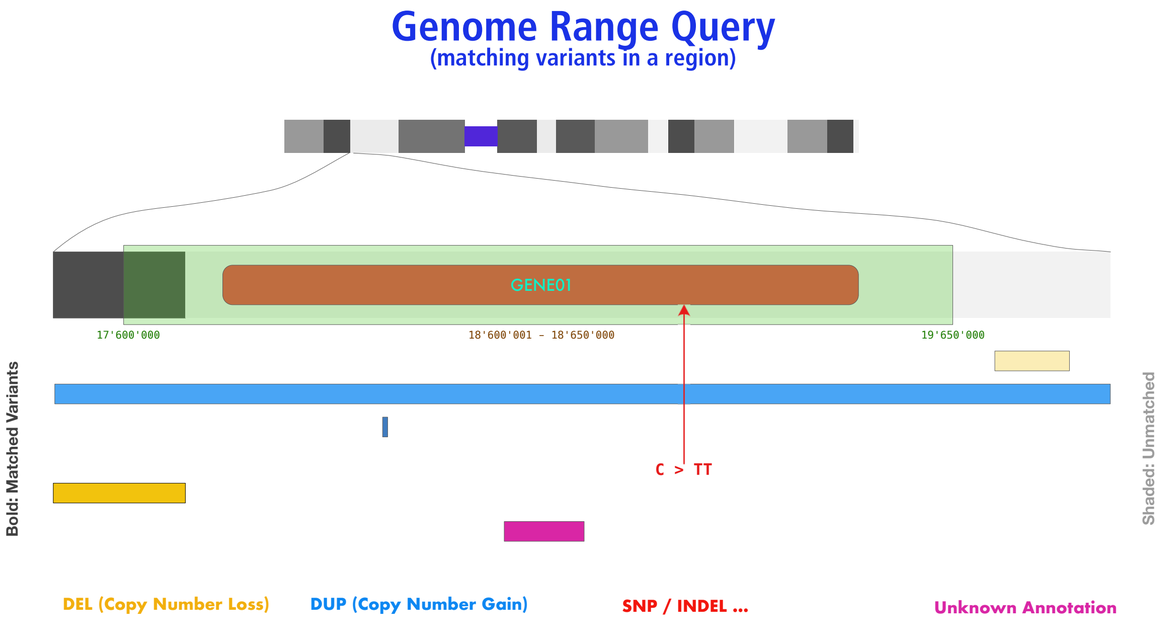
```
referenceName: "2"
start:[0,54700000]
end:[63900000,242193529]
variantType: "SO:0001742"

?referenceName=2&start=0,54700000&end=63900000,242193529&variantType=DUP
```

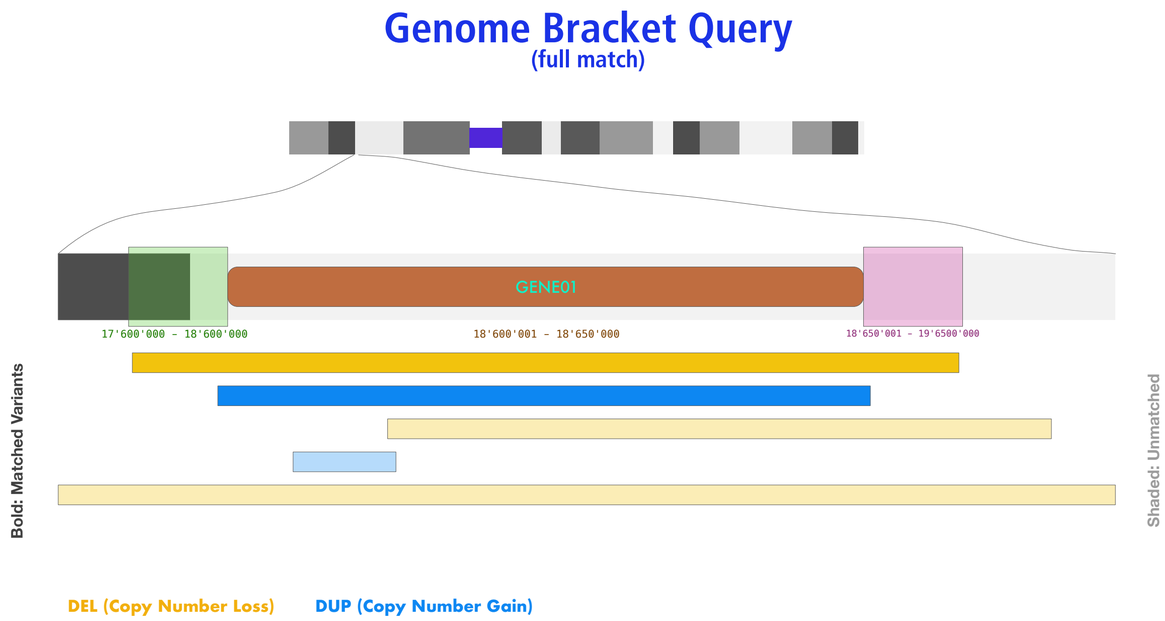
```
   Genomic region:<=----------------------|XXXXXXXXXXXX|----------------------=>
Query Range
   Start pars:    <=????????????????????????----------?????????????????????=>
   End pars:      <=----------------------??????????????????????=>
Matched variants
   Match1:        <=----------------------**********-----------------------=>
   Match2:        <=----------------------**********----------**********=>
   Match3:        <=----------------------**********-----------------------=>
   Match4:        <=--**********************************************----=>
Not Matched
   No Match1:     <=----------!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!=>
   No Match2:     <=----------!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!=>
   No Match1:     <=-------!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!=>
```

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, regerence 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

2x hCNV Implementation Studies 2021-2023

Reference hCNV datasets, use-case workflows and benchmarking

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.

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

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

hCNV Implementation Studies 2021-2023 No. 2

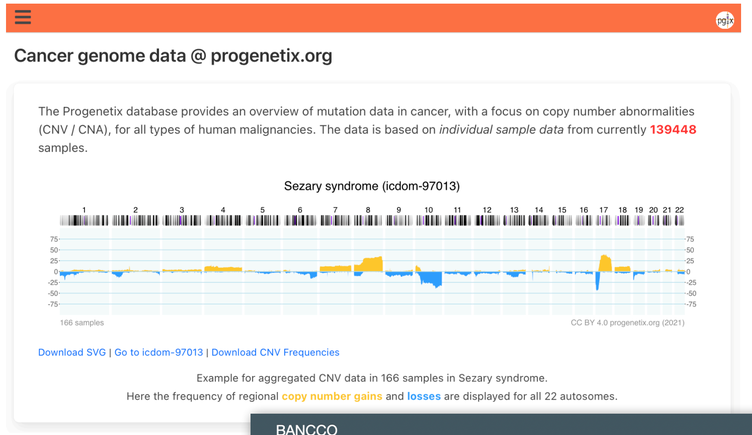
Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

- 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

hCNV Implementation Studies 2021-2023 No. 2

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



"Galaxify", "Beaconize" & "Phenopack"
Progenetix & RD-CNVdb prototypes

- 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

Beacon & Phenopackets

Data *discovery* and *delivery* using standardized GA4GH formats and schemas

- modern standards and protocols such as Beacon ***v2*** & Phenopackets ***v2*** are essential for federation and exchange of biomedical data
- emerging / established principles are the use of hierarchical coding systems and with widespread use of CURIEs
- other formats based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem

➡ Beacon query **filters** correspond well to Phenopackets data

➡ Phenopackets as supported protocol for Beacon data delivery

```
"data_use_conditions" : {
   "label" : "no restriction",
   "id" : "DUO:0000004"
},
"material" : {
   "id" : "EFO:0009656",
   "label" : "neoplastic sample"
},
{
  "ageAtDiagnosis": "P25Y3M2D"
},
"sampled_tissue" : {
     "id" : "UBERON:0002037",
     "label" : "cerebellum"
},
"histological_diagnosis" : {
     "id" : "NCIT:C3222",
     "label" : "Medulloblastoma"
},
```

Global Alliance for Genomics & Health

elixir

hCNV Implementation Studies 2021-2023

Focus on Integration with ELIXIR Platforms and Communities - and beyond

- 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

Alexander Kanitz	CH
Anthony Brookes	UK
Babita Singh	ES
Björn Grüning	DE
Christophe Béroud	FR
David Salgado	FR
Jordi Rambla	ES
Kirill Tsukanov	EMBL-EBI
Krzysztof Poterlowicz	UK
Michael Baudis	CH
Salvador Capella-Gutierrez	ES
Sergi Beltran	ES
Steven Laurie	ES
Tim Beck	UK
Timothee Cezard	EMBL-EBI