Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

SIB
Swiss Institute of
Bioinformatics

Universität
Zürich[UZH]

Reference Resources and Standards Development for Biomedical Genomics and Cancer Research

Michael Baudis

Professor of Bioinformatics
University of Zürich
Swiss Institute of Bioinformatics **SIB**
GA4GH Workstream Co-lead *DISCOVERY*
Co-lead ELIXIR Beacon API Development

Republic of Korea
Switzerland. 60th
anniversary
1963–2023

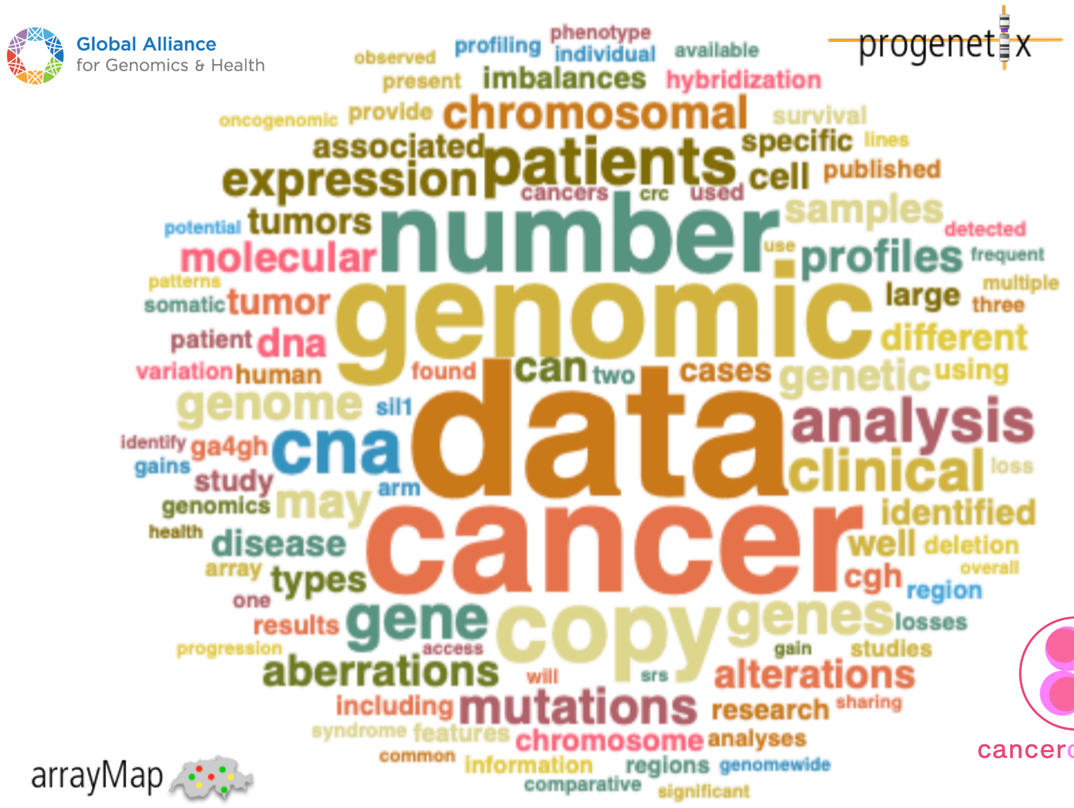
Theoretical Cytogenetics and Oncogenomics

Our work @ UZH:

- ***cancer*** genome repositories
- biocuration
- protocols & formats

Curators

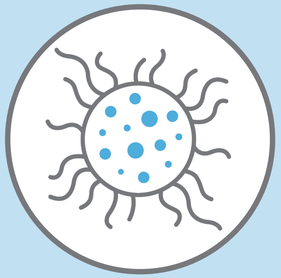
Data Parasites



Genome screening at the core of “Personalised Health”

- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts

Global Genomic Data Sharing Can...



Demonstrate patterns in health & disease

Increase statistical significance of analyses

Lead to “stronger” variant interpretations

Increase accurate diagnosis

Advance precision medicine

Different Approaches to Data Sharing

progenetix

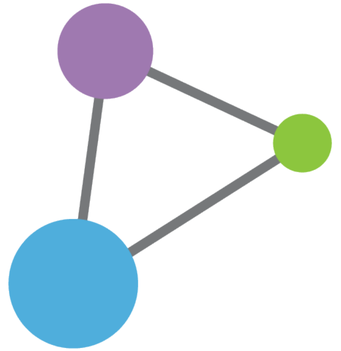
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network

Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

The Swiss Personalized Health Network (SPHN)

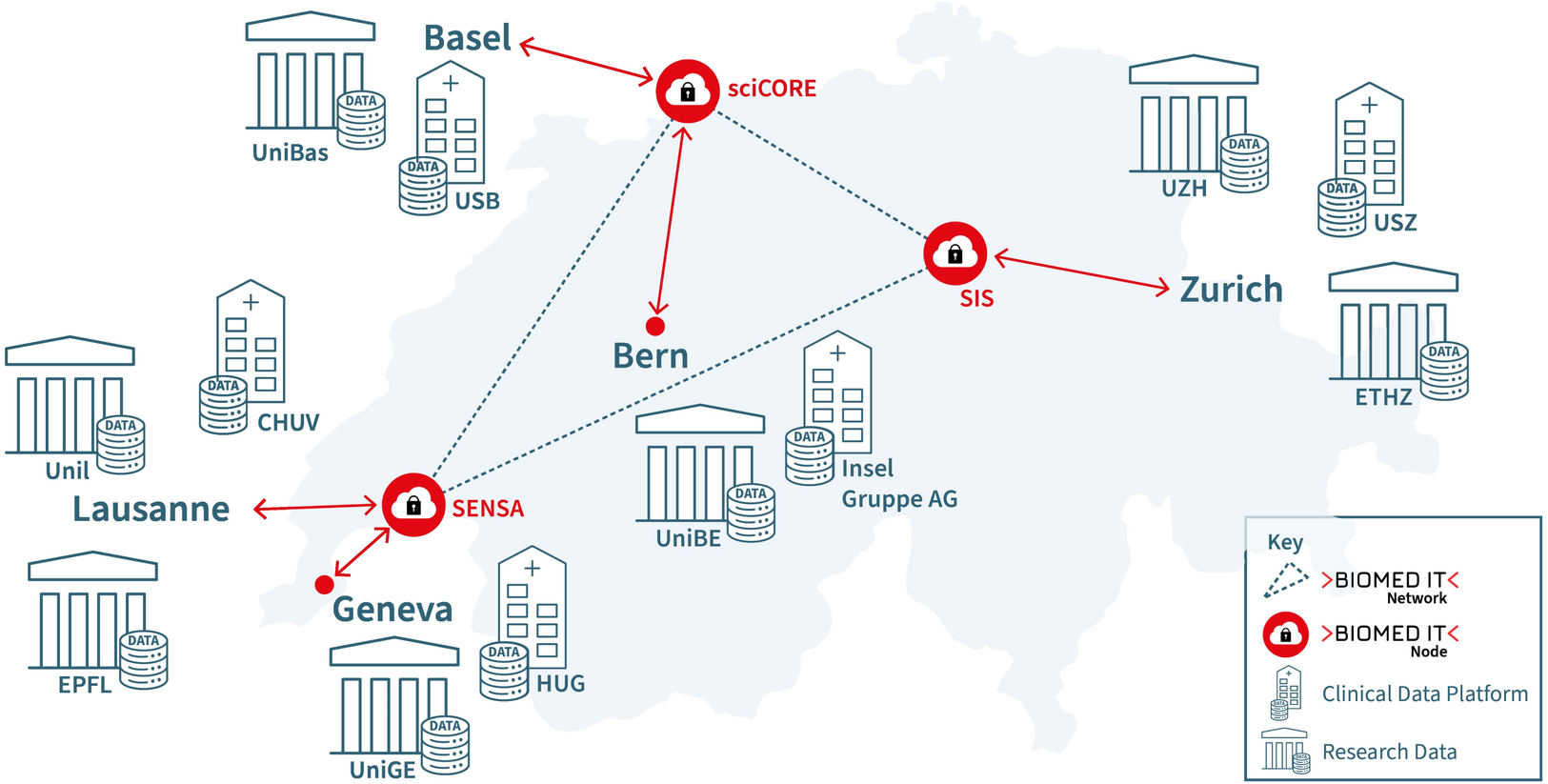
Creation of a scalable and sustainable data-enabling environment

- Including routine health data, **molecular / omics data**, registry data, clinical research data, and other health-related data types
- Research infrastructure initiative funded 2017-2024 by the Swiss Government with CHF135 million;
- Operating under a common Ethical Framework and one Information Security Policy, incl. the setup of a **Trusted Research Environment**
- Foreseen consolidation of data coordination efforts with CHF 21 million 2025-2028

→ Enable **institutions** to responsibly share interoperable health data

→ Enable **researchers** to access, integrate, and analyze data

The Swiss Personalized Health Network



Personalized Health Informatics Group

SPHN Data Coordination Center (DCC)
BioMedIT Network

Strategic Focus Area
Personalized Health and Related Technologies

Different Approaches to Data Sharing

progenetix

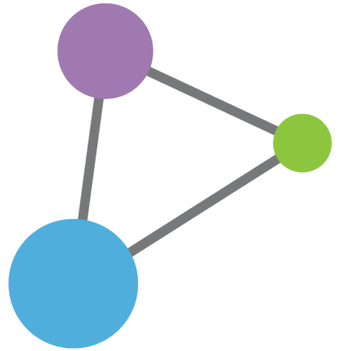
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network

Hub and Spoke
Common data elements, access, and usage rules



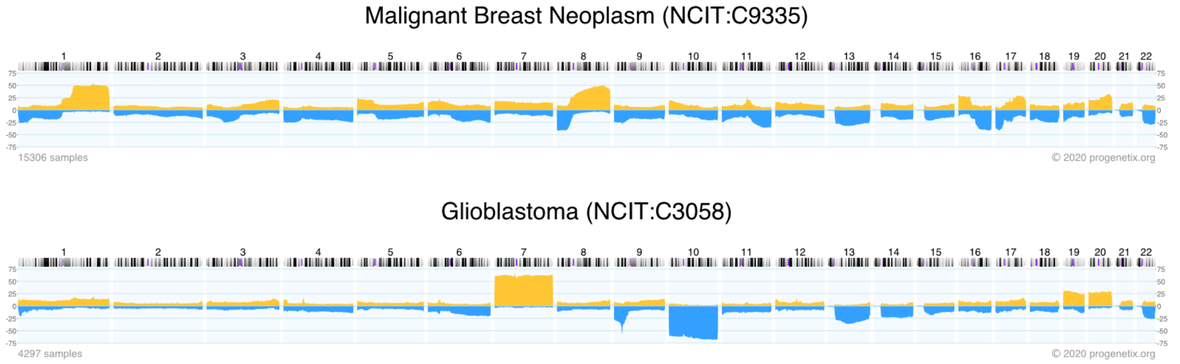
Linkage of distributed and disparate datasets

progenetix.org

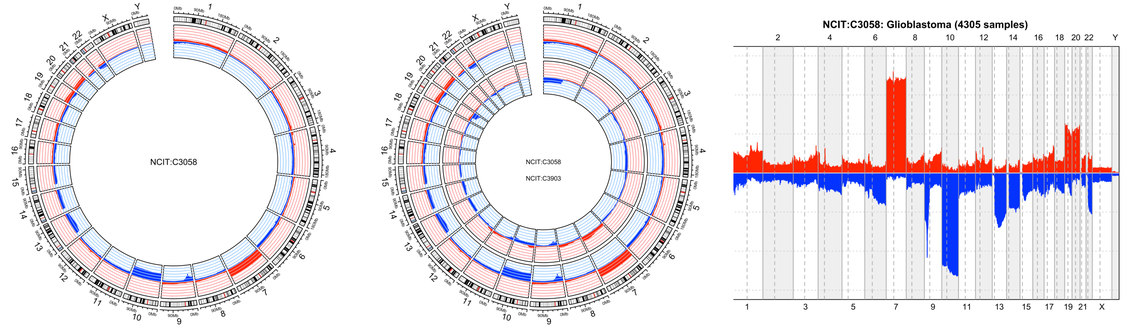
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series

Regional CNV Frequencies for >800 Cancer Types



Progenetix R API using Beacon *handover* objects

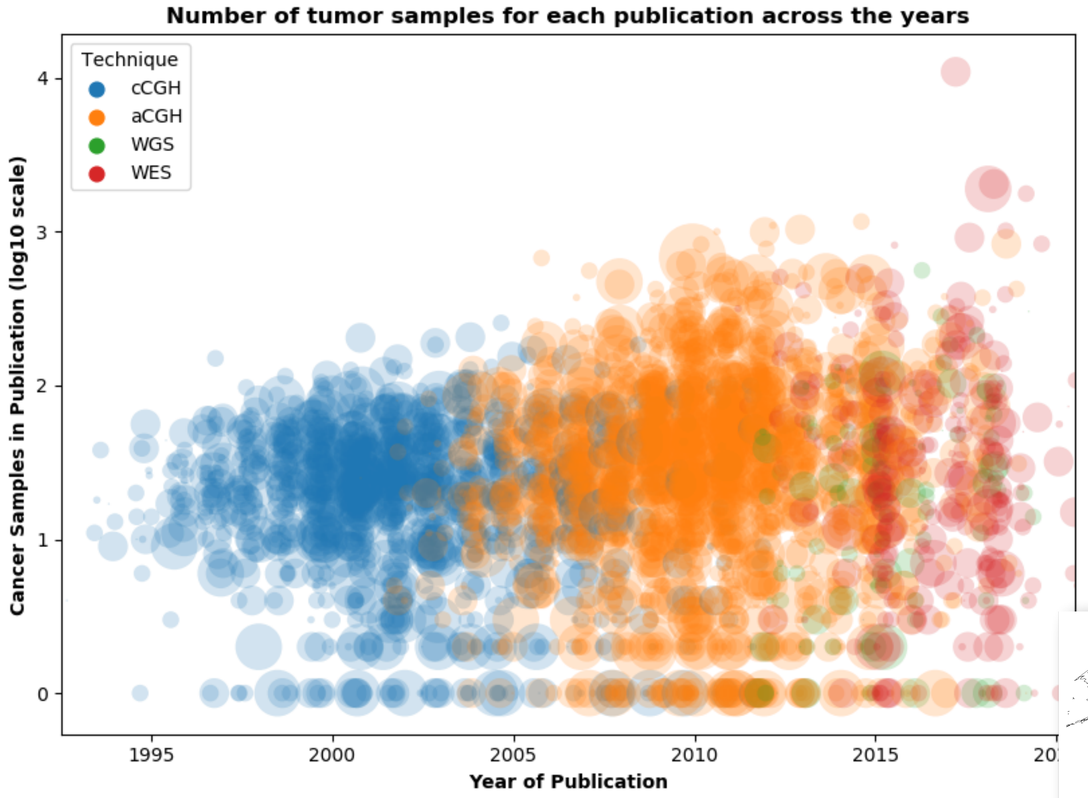


Visualization of CNV features using the pgxRpi R package. Aggregated CNV data for cancer typesis displayed using Circos or frequency plots in a local R environment.
The R package relies on the Beacon v2 API to communicate with Progenetix.

Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics



progenetix

- Cancer CNV Profiles
- Search Samples
- Studies & Cohorts
  - arrayMap
  - TCGA Samples
  - DIPG Samples
  - Gao & Baudis, 2021
  - Cancer Cell Lines
- Publication DB
- Services
  - NCIt Mappings

Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

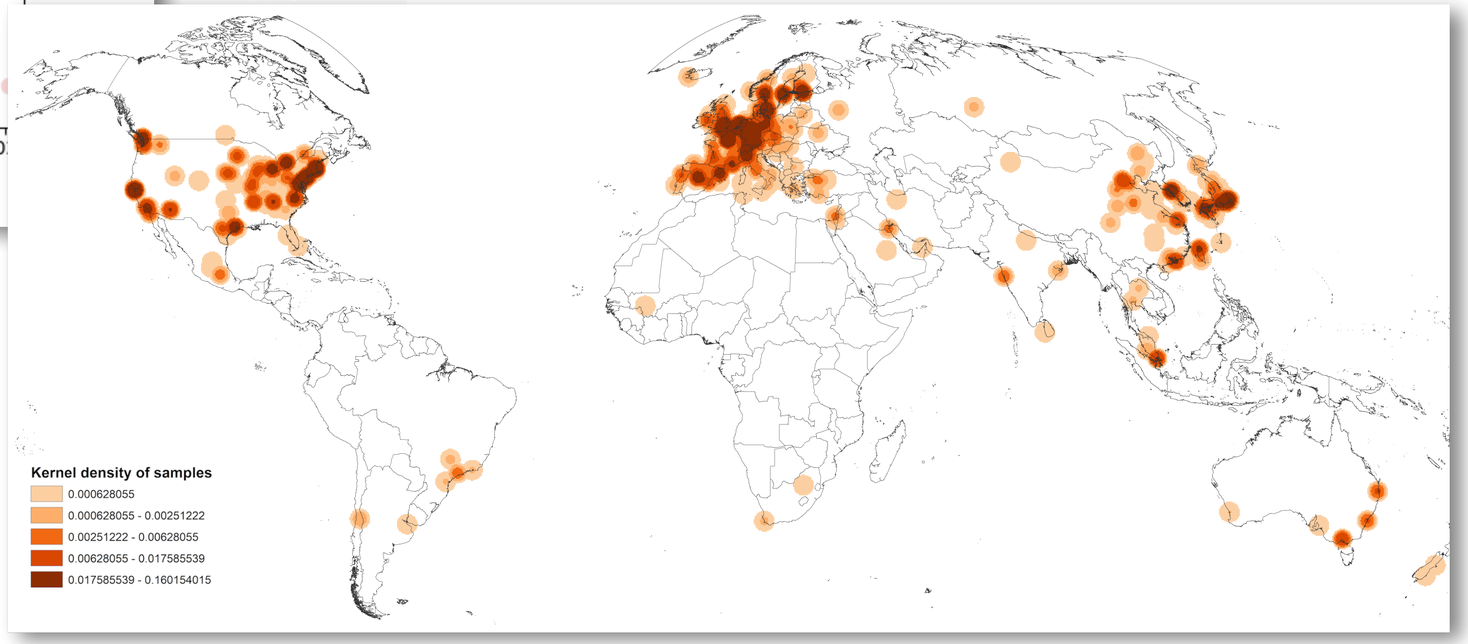
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

Filter

City

Type to search...

Publications (3324)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations	0	79	0	0	0



DATABASE
The Journal of Biological Databases and Curation

Database, 2020, 1–9
doi: 10.1093/databa/baaa009
Articles

Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo[1,2], Elise Acheson[3], Qingyao Huang[1,2] and Michael Baudis[1,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [2]Swiss Institute of Bioinformatics, Zurich, Switzerland [3]Department of Geography, University of Zurich, Zurich, Switzerland

Map of the geographic distribution (by affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets. The numbers are derived from the 3'240 publications in Progenetix.

Database, 2021, **2021(0)**, 1–9
DOI: https://doi.org/10.1093/database/baab043
Database update

The Progenetix oncogenomic resource in 2021

Qingyao Huang[1,2], Paula Carrio-Cordo[1,2], Bo Gao[1,2], Rahel Paloots[1,2] and Michael Baudis[1,2,*]

[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
[2]Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland

*Corresponding author: Tel: +41 44 635 34 86; Email: michael.baudis@mls.uzh.ch

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. *et al.* The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: https://doi.org/10.1093/database/baab043

Abstract

In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identify and characterize molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 138 663 (including 115 357 tumor) copy number variation (CNV) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNV landscape summary over 51 distinctive National Cancer Institute Thesaurus cancer terms as well as updates in database schemas, and data access including new web front-end and programmatic data access.

Database URL: progenetix.org

Table 1. Statistics of samples from various data resources

Data source	GEO	ArrayExpress	cBioPortal	TCGA	Total
No. of studies	898	51	38	33	1939
No. of samples	63 568	4351	19 712	22 142	138 663
Tumor	52 090	3887	19 712	11 090	115 357
Normal	11 478	464	0	11 052	23 306
Classifications					
ICD-O (Topography)	100	54	88	157	209
ICD-O (Morphology)	246	908	265	140	491
NCIt	346	148	422	182	788
Collections					
Individuals	63 568	4351	19 712	10 995	127 549
Biosamples	63 568	4351	19 712	22 142	138 663
Callsets[a]	63 568	4351	19 712	22 376	138 930
Variants	5 514 126	118 4170	1 778 096	2 654 065	10 716 093

[a]set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

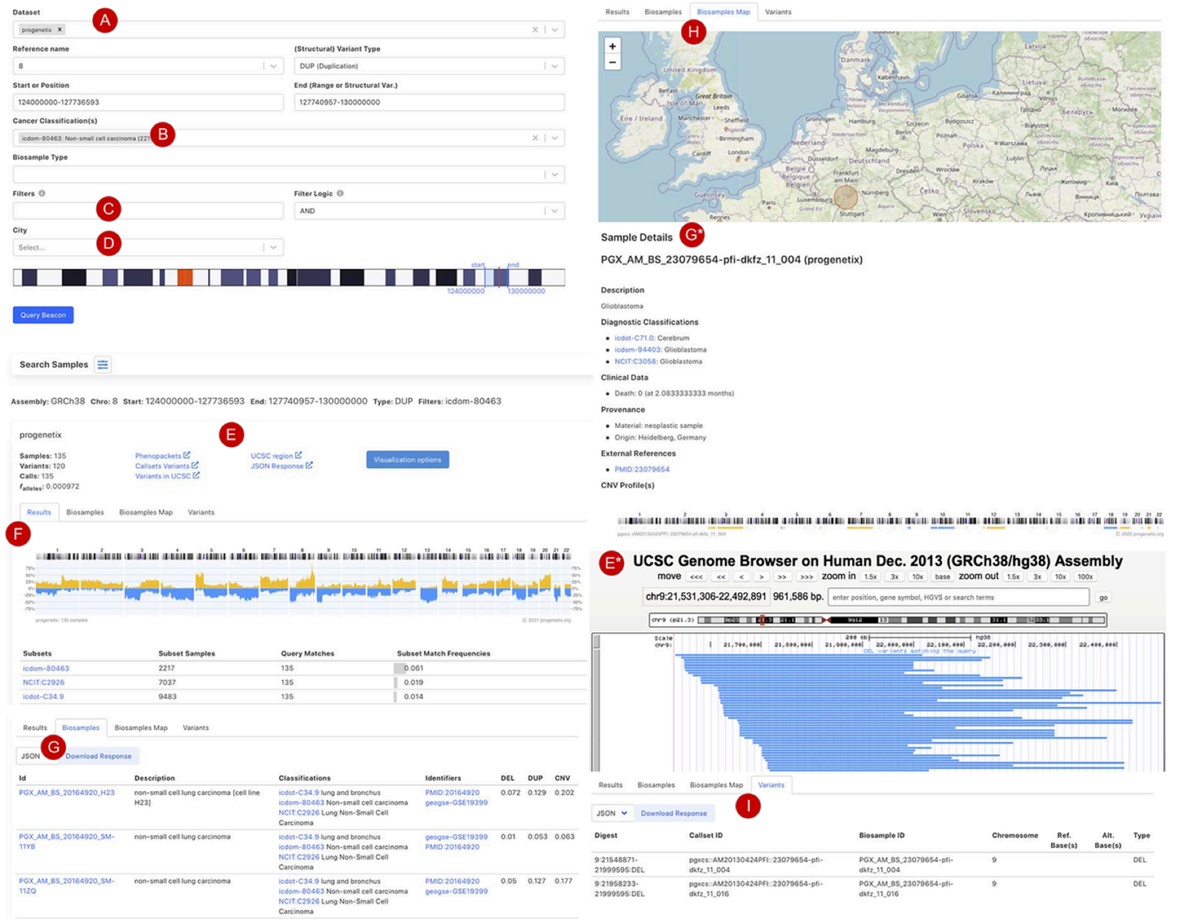


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellosaurus; D: additional filter for geographic location; E: external link to UCSC browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched biosamples with description, statistics and reference. More detailed biosample information can be viewed through 'id' link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.

cancercelllines.org

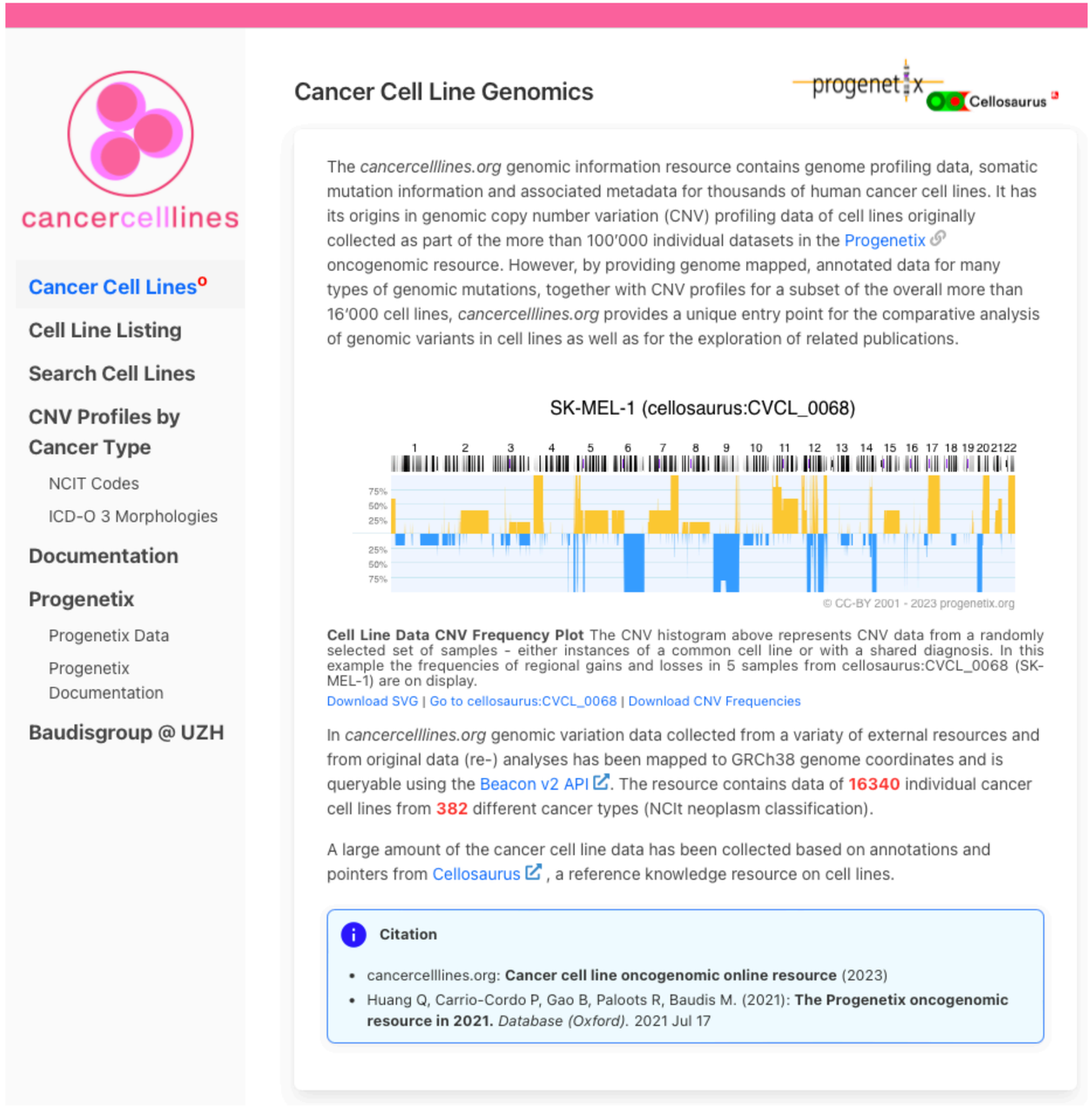
Cancer Cell Line Genomics Resource

- cancercelllines.org built on Progenetix platform
- includes over **5600 cell line CNV profiles**
- cancer cell line variants ()SNV, INDELs ...) for **16178 cell lines** from 400 different disease classifications
- mapped to to *Cellosaurus*
- hierarchical representation ("derived from"...)
- SNVs mapped from ClinVar with variant severity and disease ontologies
- CCLE per cell line include variant effect
- CNV profiles allow temporal stability estimates and tumor type similarity matching

Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics



Maintaining some Standards

CNV Term Use in Computational (File/Schema) Formats

https://cnvar.org/resources/CNV-annotation-standards/

EFO	Beacon	VCF	SO	GA4GH VRS ⇒ VRS proposal[1]	Notes
EFO:0030070 copy number gain	DUP [2] or EFO:0030070	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	low-level gain (implicit) ⇒ EFO:0030070 copy number gain	a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence
EFO:0030071 low-level copy number gain	DUP [2] or EFO:0030071	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	low-level gain ⇒ EFO:0030071 low-level copy number gain	
EFO:0030072 high-level copy number gain	DUP [2] or EFO:0030072	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030072 high-level copy number gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
EFO:0030073 focal genome amplification	DUP [2] or EFO:0030073	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030073 focal genome amplification	commonly but not consistently used for >=5 copies on a bi-allelic genome region, of limited size (operationally max. 1-5Mb)
EFO:0030067 copy number loss	DEL [2] or EFO:0030067	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	partial loss (implicit) ⇒ EFO:0030067 copy number loss	a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence
EFO:0030068 low-level copy number loss	DEL [2] or EFO:0030068	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0030068 low-level copy number loss	
EFO:0020073 high-level copy number loss	DEL [2] or EFO:0020073	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0020073 high-level copy number loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted
EFO:0030069 complete genomic deletion	DEL [2] or EFO:0030069	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	complete loss ⇒ EFO:0030069 complete genomic deletion	complete genomic deletion (e.g. homozygous deletion on a bi-allelic genome region)

Hangjia Zhao
Michael Baudis
(& the VRS group!)

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH Standards for Federated Genomic Data Discovery

Different Approaches to Data Sharing

progenetix

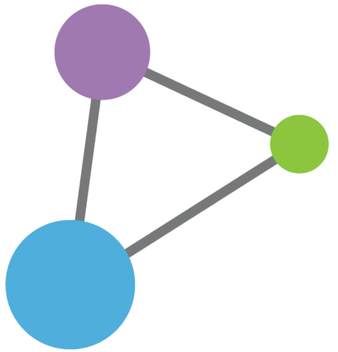
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network

Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

INFORMATICS

Beacon v2 and Beacon networks: A "lingua federated data discovery n biomedical ger

Jordi Rambla[1,2] | Michael Baudis[3] | Roberto Ariosa[1] | Tim Beck[4] | Lauren A. Fromont[1] | Arcadi Navarro[1,5,6,7] | Rahel Paloots[3] | Manuel Rueda[1] | Gary Saunders[8] | Babita Singh[1] | John D. Spalding[9] | Juha Törnroos[9] | Claudia Vasallo[1] | Colin D. Veal[4] | Anthony J. Brookes[4]

Cell Genomics

CellPress
OPEN ACCESS

Commentary

International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,[1,2,*] Heidi L. Rehm,[3,4] Peter Goodhand,[5,6] Angela J.H. Page,[4,5] Yann Joly,[2] Michael Baudis,[7] Jordi Rambla,[8,9] Arcadi Navarro,[8,10,11,12] Tommi H. Nyronen,[13,14] Mikael Linden,[13,14] Edward S. Dove,[15] Marc Fiume,[16] Michael Brudno,[17] Melissa S. Cline,[18] and Ewan Birney[19]

Cell Genomics

CellPress
OPEN ACCESS

Perspective

GA4GH: International policies and standards for data sharing across genomic research and healthcare

Heidi L. Rehm,[1,2,47] Angela J.H. Page,[1,3,*] Lindsay Smith,[3,4] Jeremy B. Adams,[3,4] Gil Alterovitz,[5,47] Lawrence J. Babb,[1] Maxmillian P. Barkley,[6] Michael Baudis,[7,8] Michael J.S. Beauvais,[3,9] Tim Beck,[10] Jacques S. Beckmann,[11] Sergi Beltran,[12,13,14] David Bernick,[1] Alexander Bernier,[9] James K. Bonfield,[15] Tiffany F. Boughtwood,[16,17] Guillaume Bourque,[9,18] Sarion R. Bowers,[15] Anthony J. Brookes,[10] Michael Brudno,[18,19,20,21,38] Matthew H. Brush,[22] David Bujold,[9,18,38] Tony Burdett,[23] Orion J. Buske,[24] Moran N. Cabili,[1] Daniel L. Cameron,[25,26] Robert J. Carroll,[27] Esmeralda Casas-Silva,[123] Debyani Chakravarty,[29] Bimal P. Chaudhari,[30,31] Shu Hui Chen,[32] J. Michael Cherry,[33] Justina Chung,[3,4] Melissa Cline,[34] Hayley L. Clissold,[15] Robert M. Cook-Deegan,[35] Mélanie Courtot,[23] Fiona Cunningham,[23] Miro Cupak,[6] Robert M. Davies,[15] Danielle Denisko,[19] Megan J. Doerr,[36] Lena I. Dolman,[19]

(Author list continued on next page)

Cell Genomics

CellPre
OPEN ACC

Technology

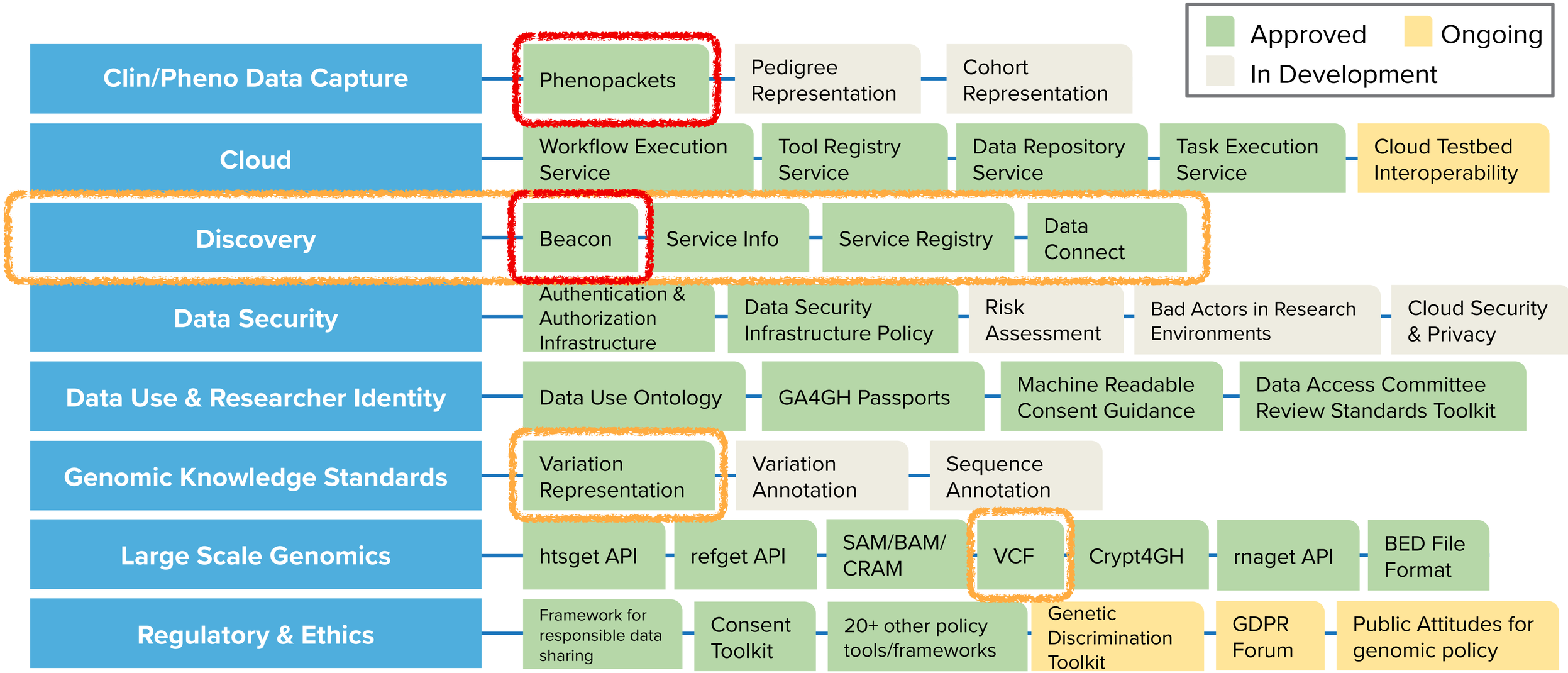
The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification

Alex H. Wagner,[1,2,25,*] Lawrence Babb,[3,*] Gil Alterovitz,[4,5] Michael Baudis,[6] Matthew Brush,[7] Daniel L. Cameron,[8,9] Melissa Cline,[10] Malachi Griffith,[11] Obi L. Griffith,[11] Sarah E. Hunt,[12] David Kreda,[13] Jennifer M. Lee,[14] Stephanie Li,[15] Javier Lopez,[16] Eric Moyer,[17] Tristan Nelson,[18] Ronak Y. Patel,[19] Kevin Riehle,[19] Peter N. Robinson,[20] Shawn Rynearson,[21] Helen Schuilenburg,[12] Kirill Tsukanov,[12] Brian Walsh,[7] Melissa Konopko,[15] Heidi L. Rehm,[3,22] Andrew D. Yates,[12] Robert R. Freimuth,[23] and Reece K. Hart[3,24,*]

The GA4GH Phenopacket schema defines a computable representation of clinical data

NATURE BIOTECHNOLOGY | VOL 40 | JUNE 2022 | 815–820

Overview of GA4GH standards and frameworks



Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

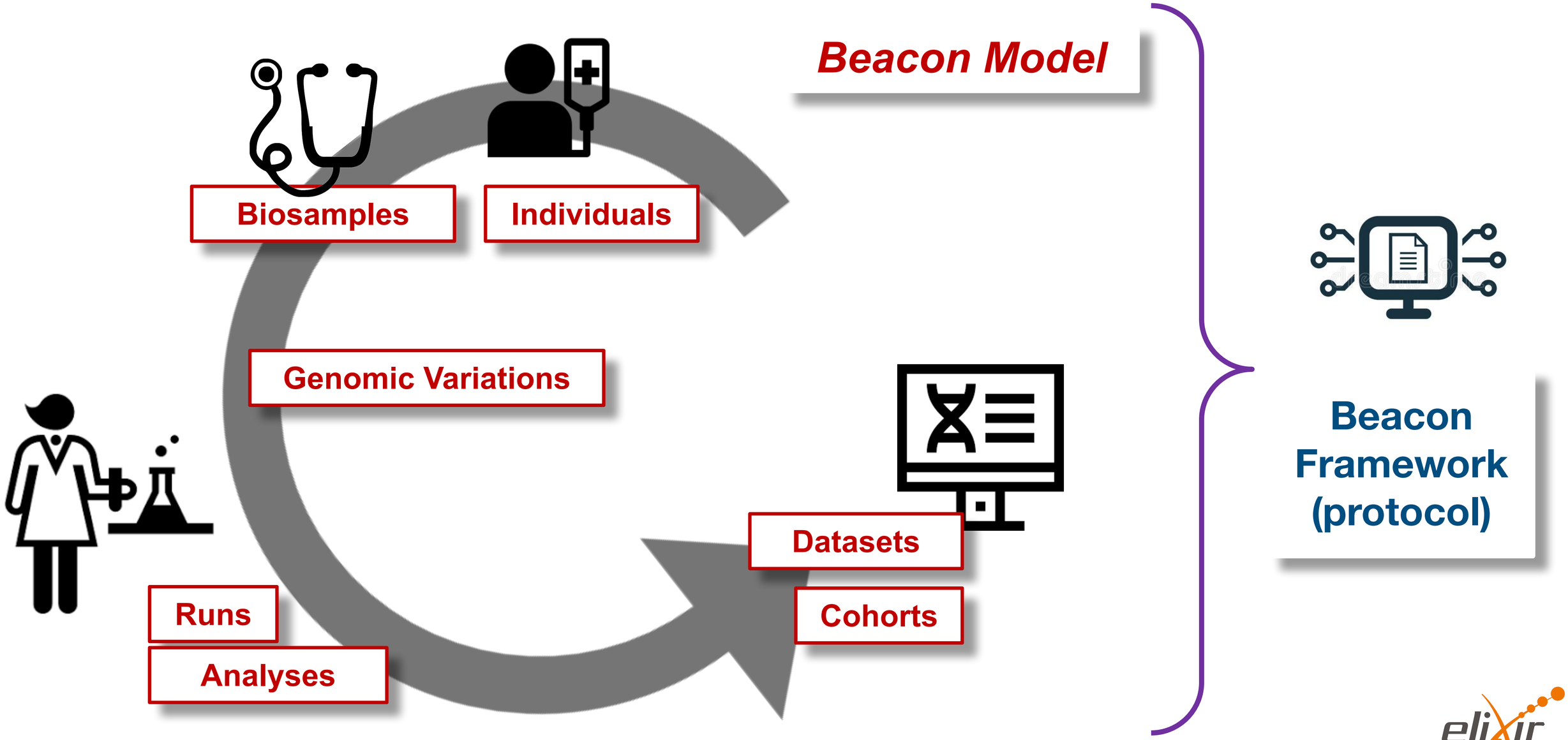
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	**GA4GH founding event; Jim Ostell proposes Beacon concept with "more features… version 2"**		
2015	• beacon-network.org aggregator created by DNAstack • Beacon v0.3 release	• Beacon+ concept implemented on progenetix.org	• ELIXIR starts Beacon project support
2016	• work on queries for structural variants (brackets for fuzzy start and end parameters...)	• concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating CNV parameters (e.g. "startMin, statMax")	• Beacon+ demos "handover" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • GA4GH Beacon v1 approved at Oct plenary	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off	• new Beacon website (March)
2019	• ELIXIR Beacon Network	• adopting "handover" concept • "Scouts" teams working on different aspects - filters, genomic variants, compliance ... • discussions w/ clinical stakeholders	• Beacon publication at Nature Biotechnology
2020		• framework + models concept implemented	
2021		• range and bracket queries, variant length • starting of GA4GH review process	
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved at Apr GA4GH Connect**	• *docs.genomebeacons.org*

Beacon v2

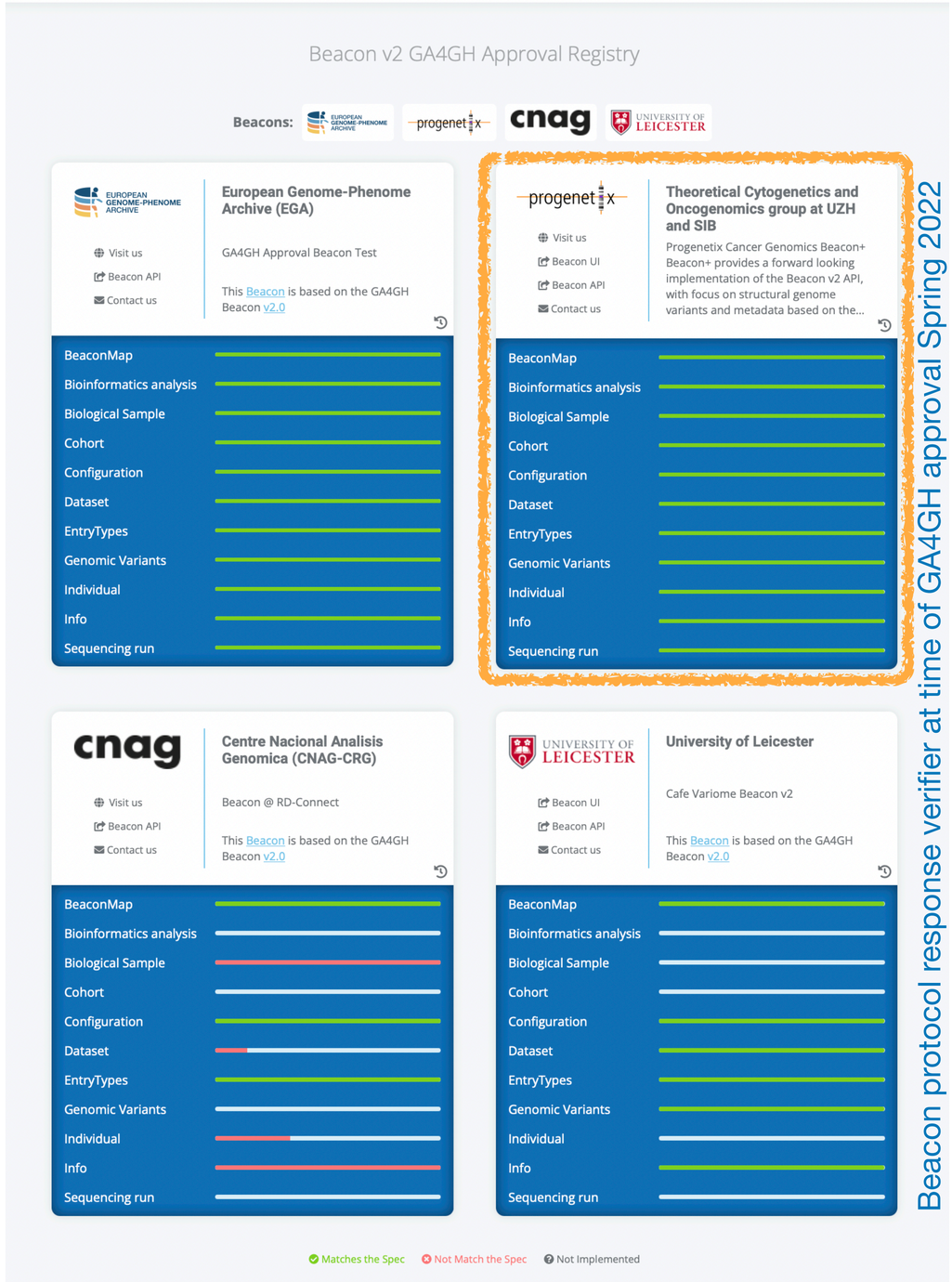
docs.genomebeacons.org



Progenetix & Beacon

Implementation driven standards development

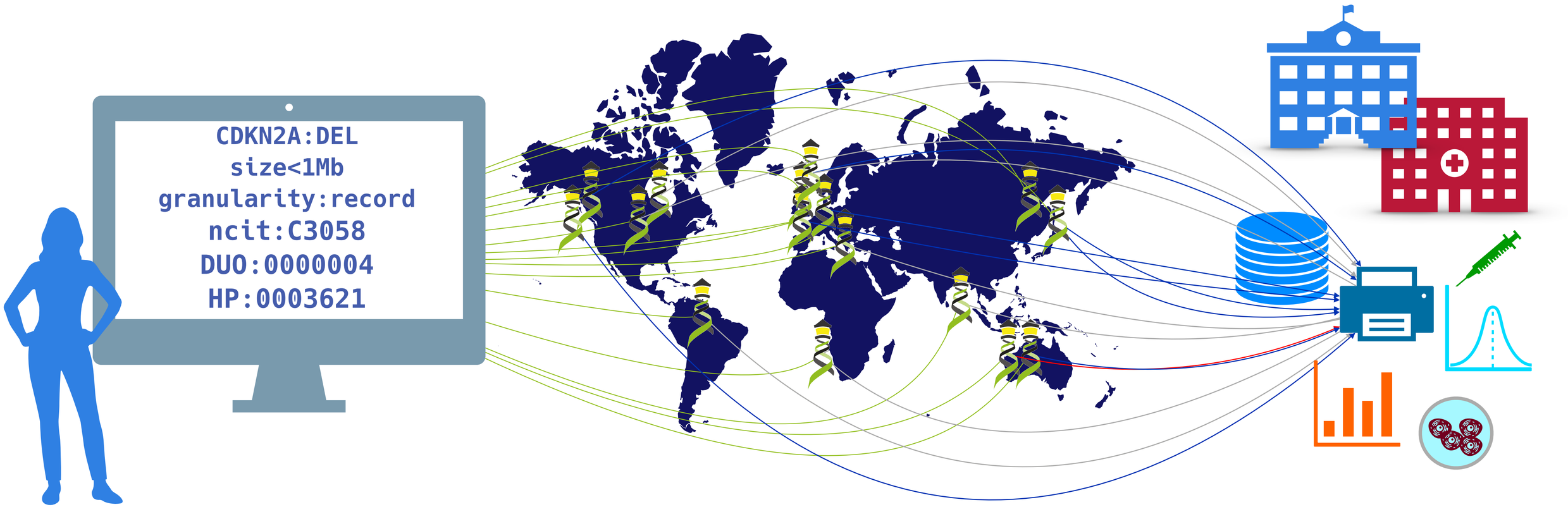
- Progenetix Beacon+ has served as implementation driver since 2016
- prototyping of advanced Beacon features such as
  - ➡ structural variant queries
  - ➡ data handovers
  - ➡ Phenopackets integration



Global Alliance for Genomics & Health

progenetix

elixir



Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

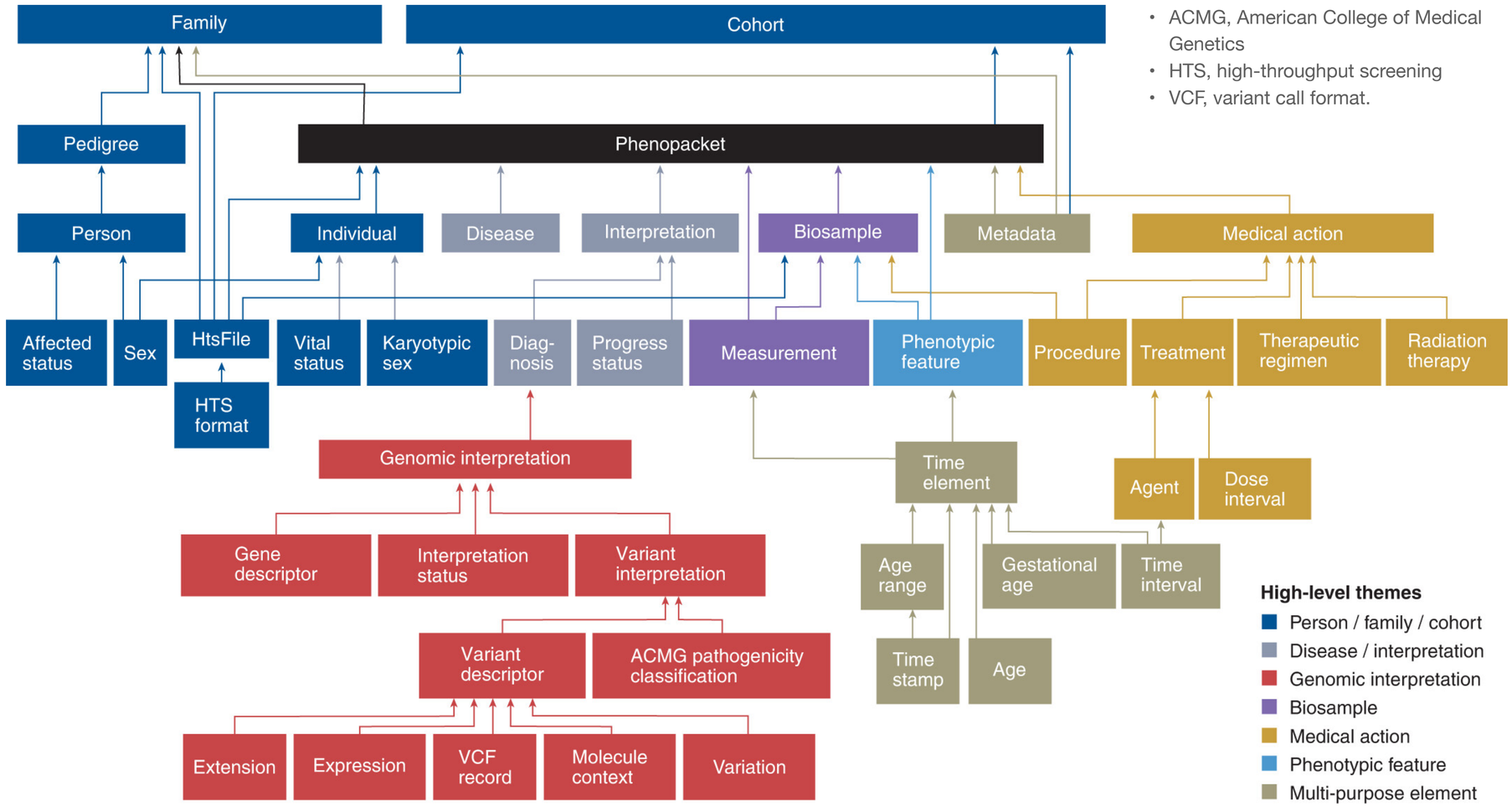
Beacon *v2* API

The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval

The GA4GH Phenopackets v2 Standard
A Computable Representation of Clinical Data

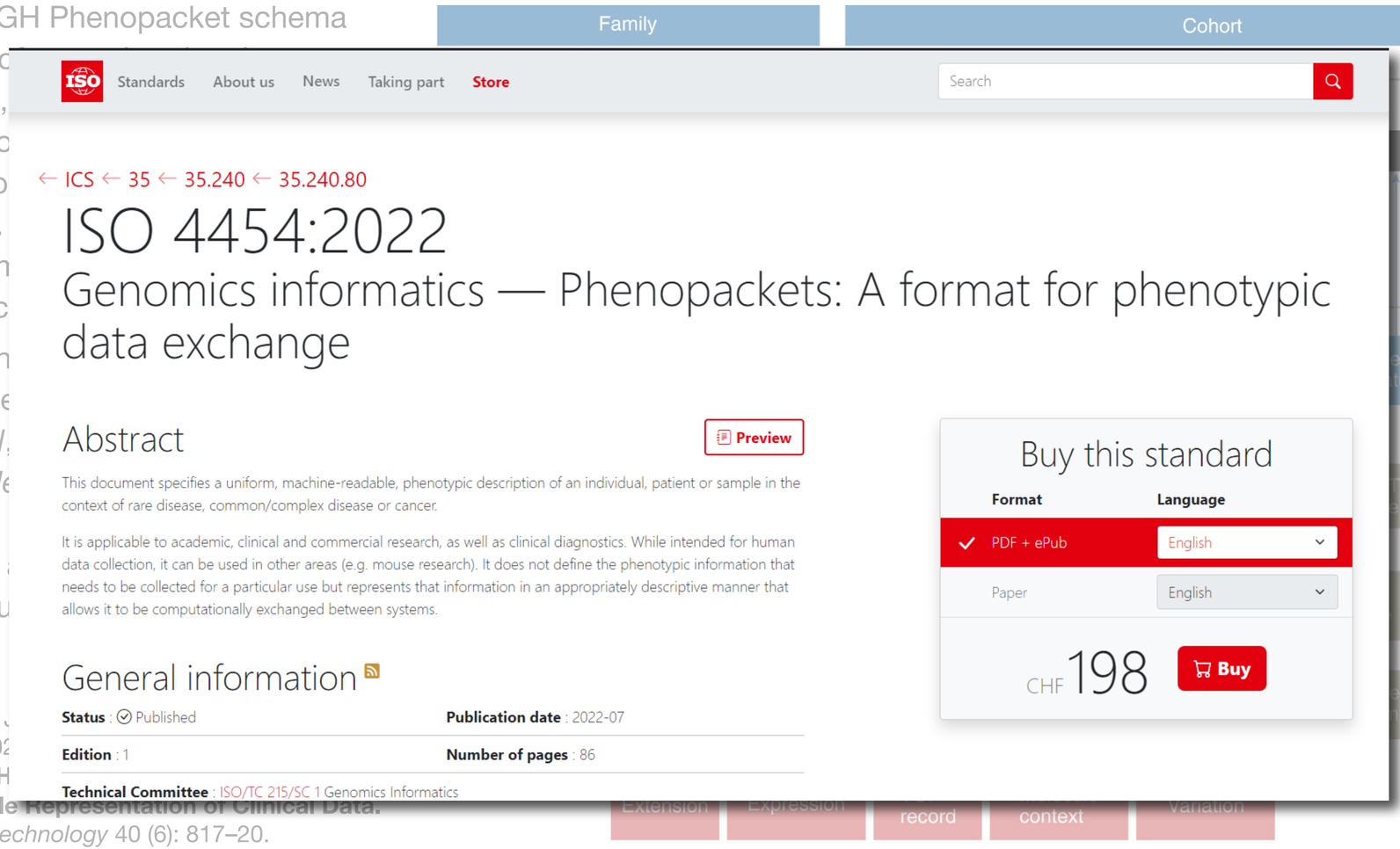
The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain topic, such as phenotype, variant or pedigree. An element can contain other elements, which allows a hierarchical representation of data.

For instance, Phenopacket contains elements of type *Individual*, *PhenotypicFeature*, *Biosample* and so on. Individual elements can therefore be regarded as **building blocks** of larger structures.



Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, *et al.* 2022.
The GA4GH Phenopacket Schema Defines a Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.

The GA4GH Phenopackets v2 Standard
A Computable Representation of Clinical Data



- ACMG, American College of Medical Genetics
- HTS, high-throughput screening
- VCF, variant call format.

Co-hosted by Smart Health Standards Forum, ISO TC215 SC1 Korean mirror committee

The GA4GH Phenopacket schema consists o
elements,
informatic
such as p
pedigree.
other elen
hierarchic

For instan
contains e
Individual,
Biosample
elements
regarded
larger stru

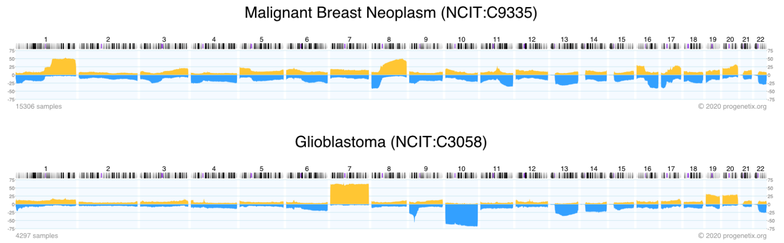
Jacobsen
Beckmann
TJ, *et al.* 202
The GA4GH
Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.

Theoretical Cytogenetics and Oncogenomics @baudisgroup

Swiss Institute of Bioinformatics

Universität Zürich[UZH]

- **curated resources**, patterns & markers in cancer genomics, especially somatic **structural genome variants**
- bioinformatics in **collaborative studies**
- bioinformatics **tools** & methods
- **standards** and implementations for **data sharing** in genomics, personalized health
- open research data "**ambassadoring**"



progenetix

cancercelllines

The Beacon team through the ages

Jordi Rambla
Arcadi Navarro
Roberto Ariosa
Manuel Rueda
Lauren Fromont
Mauricio Moldes
Claudia Vasallo
Babita Singh
Sabela de la Torre
Marta Ferri
Fred Haziza

Juha Törnroos
Teemu Kataja
Ilkka Lappalainen
Dylan Spalding

Tony Brookes
Tim Beck
Colin Veal
Tom Shorter

Michael Baudis
Rahel Paloots
Hangjia Zhao
Ziying Yang
Bo Gao

Augusto Rendon
Ignacio Medina
Javier López
Jacobo Coll
Antonio Rueda

Sergi Beltran
Carles Hernandez

David Salgado

Salvador Capella
Dmitry Repchevski
JM Fernández

Laura Furlong
Janet Piñero

Serena Scollen
Gary Saunders
Giselle Kerry
David Lloyd

Nicola Mulder
Mamana Mbiyavanga
Ziyaad Parker

David Torrents

Dean Hartley

Joaquin Dopazo
Javier Pérez
J.L. Fernández
Gema Roldan

Thomas Keane
Melanie Courtot
Jonathan Dursi

Heidi Rehm
Ben Hutton

Toshiaki Katayama

Stephane Dyke

Marc Fiume
Miro Cupak

Melissa Cline

Diana Lemos

GA4GH Phenopackets
Peter Robinson
Jules Jacobsen

GA4GH VRS
Alex Wagner
Reece Hart

Beacon PRC
Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga

Alice Mann
Neerjah Skantharajah

docs.genomebeacons.org


Universität Zürich[UZH]
SIB
Swiss Institute of Bioinformatics
Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.
elixir
SPHN