



University of  
Zurich<sup>UZH</sup>

# BIO392 Bioinformatics of Genome Variations

Survival | Classifications

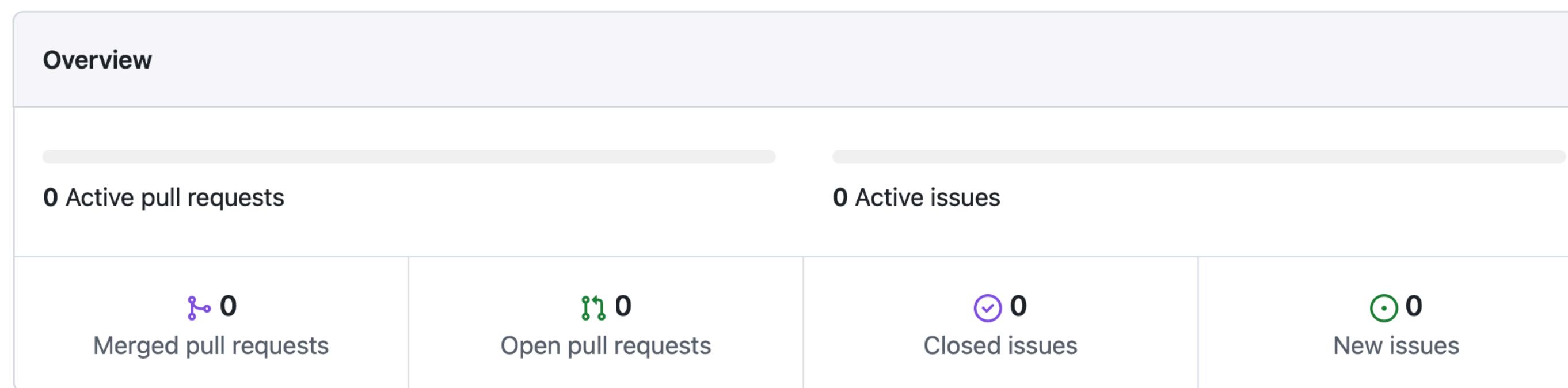
Michael Baudis **UZH SIB**  
Computational Oncogenomics

# BIO392 FS25

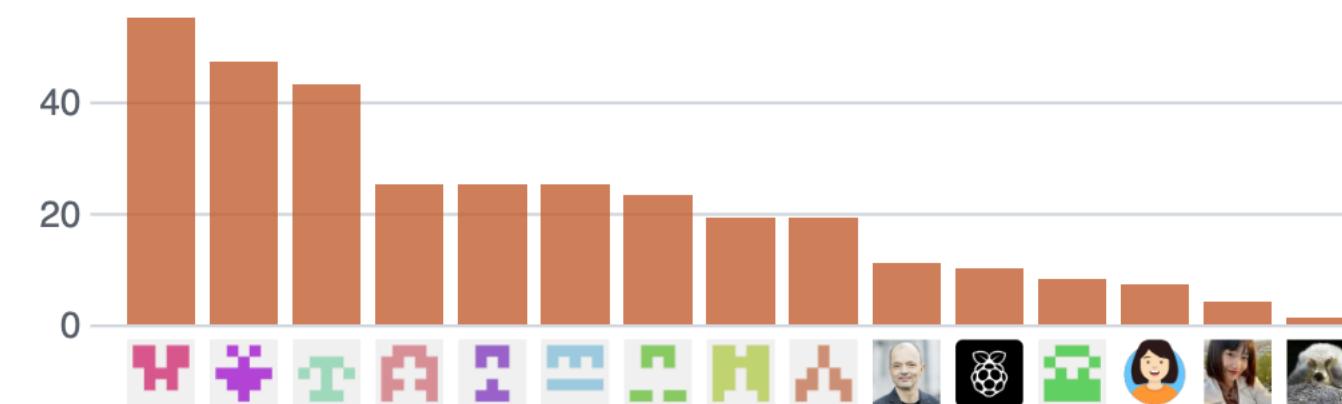
## Github Activity

March 16, 2025 – April 16, 2025

Period: 1 month ▾



Excluding merges, **15 authors** have pushed **322 commits** to master and **322 commits** to all branches. On master, **124 files** have changed and there have been **6,488 additions** and **2 deletions**.



```
→ ~ pip3 install mkdocs  
→ ~ cd ~/Github/UZH-BI0392  
→ UZH-BI0392 git: (master) mkdocs serve
```

```
→ ΟΖΗ-ΒΙ0392 git: (master) mkdocs serve
```

# **Survival**

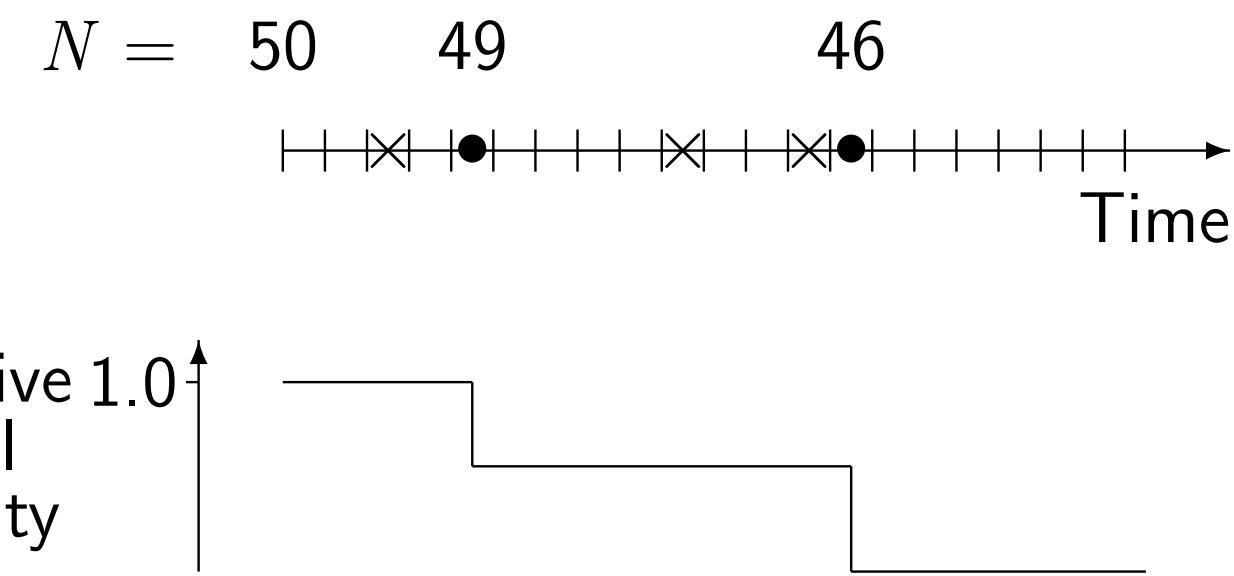
**Kaplan-Meier Analysis of Survival Based on Conditional Probabilities**

# The Kaplan-Meier Method

- ▶ The most common method of estimating the survival function.
  - ▶ A non-parametric method.
  - ▶ Divides time into small intervals where the intervals are defined by the unique times of failure (death).
  - ▶ Based on conditional probabilities as we are interested in the probability a subject surviving the next time interval given that they have survived so far.

# Kaplan–Meier method illustrated

(• = failure and × = censored):



- ▶ Steps caused by multiplying by  $(1 - 1/49)$  and  $(1 - 1/46)$  respectively
  - ▶ Late entry can also be dealt with

## Kaplan-Meier estimators (km-na)

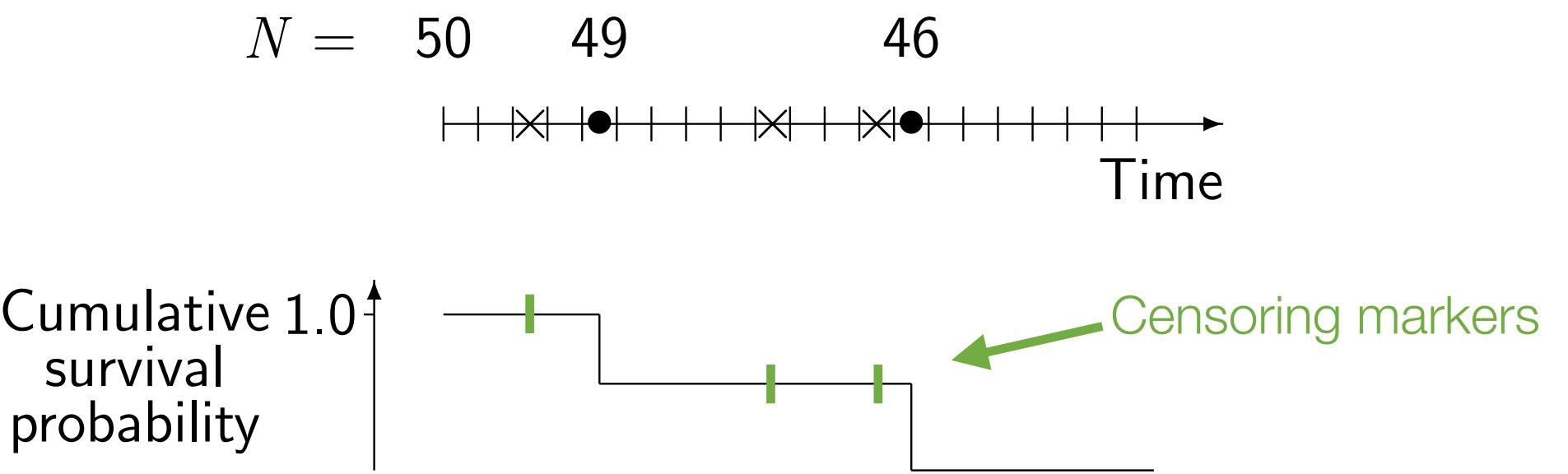
## Kaplan-Meier estimators (km-na)

## The Kaplan-Meier Method

- ▶ The most common method of estimating the survival function.
- ▶ A non-parametric method.
- ▶ Divides time into small intervals where the intervals are defined by the unique times of failure (death).
- ▶ Based on conditional probabilities as we are interested in the probability a subject surviving the next time interval given that they have survived so far.

## Kaplan–Meier method illustrated

(• = failure and × = censored):



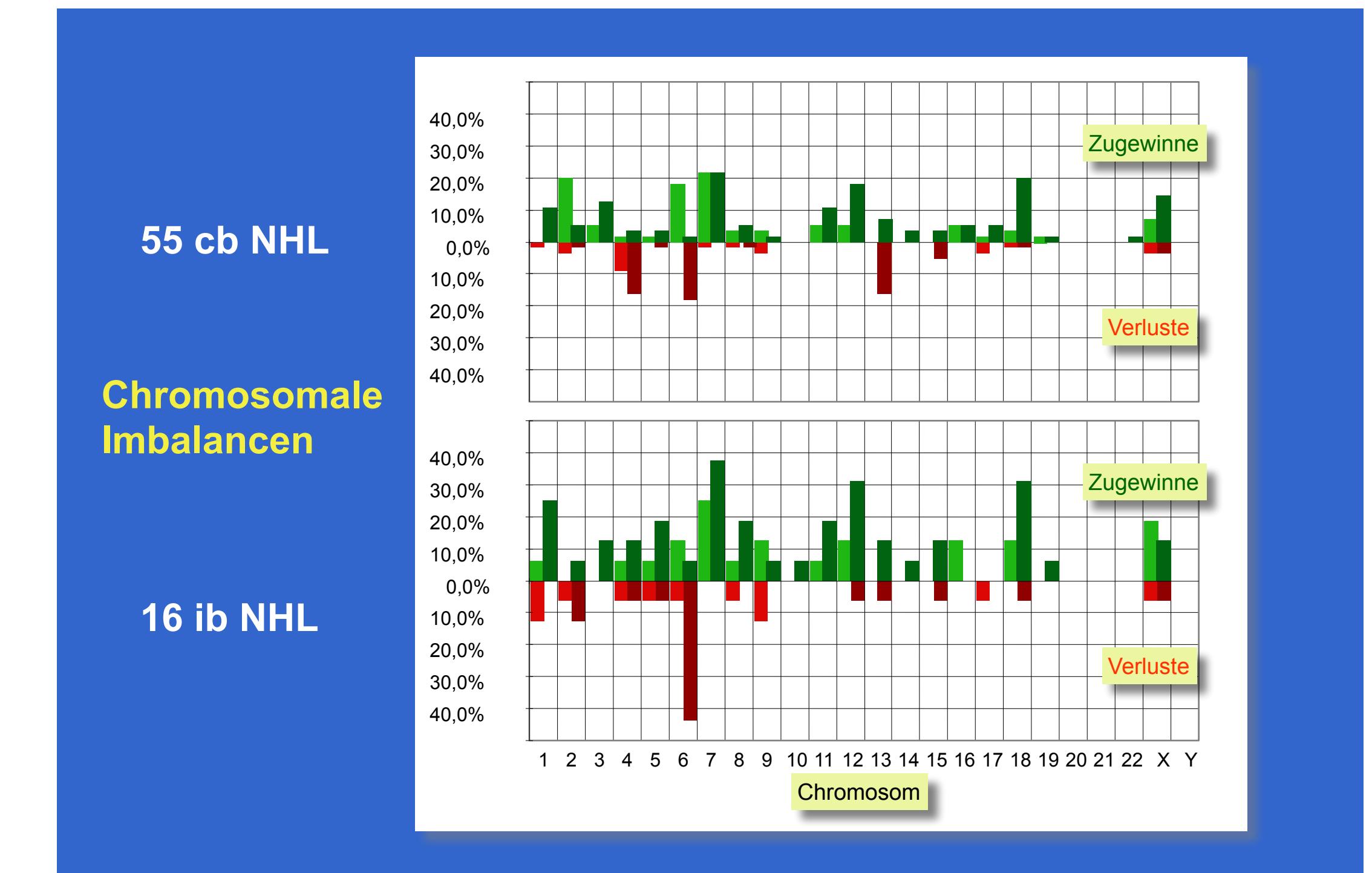
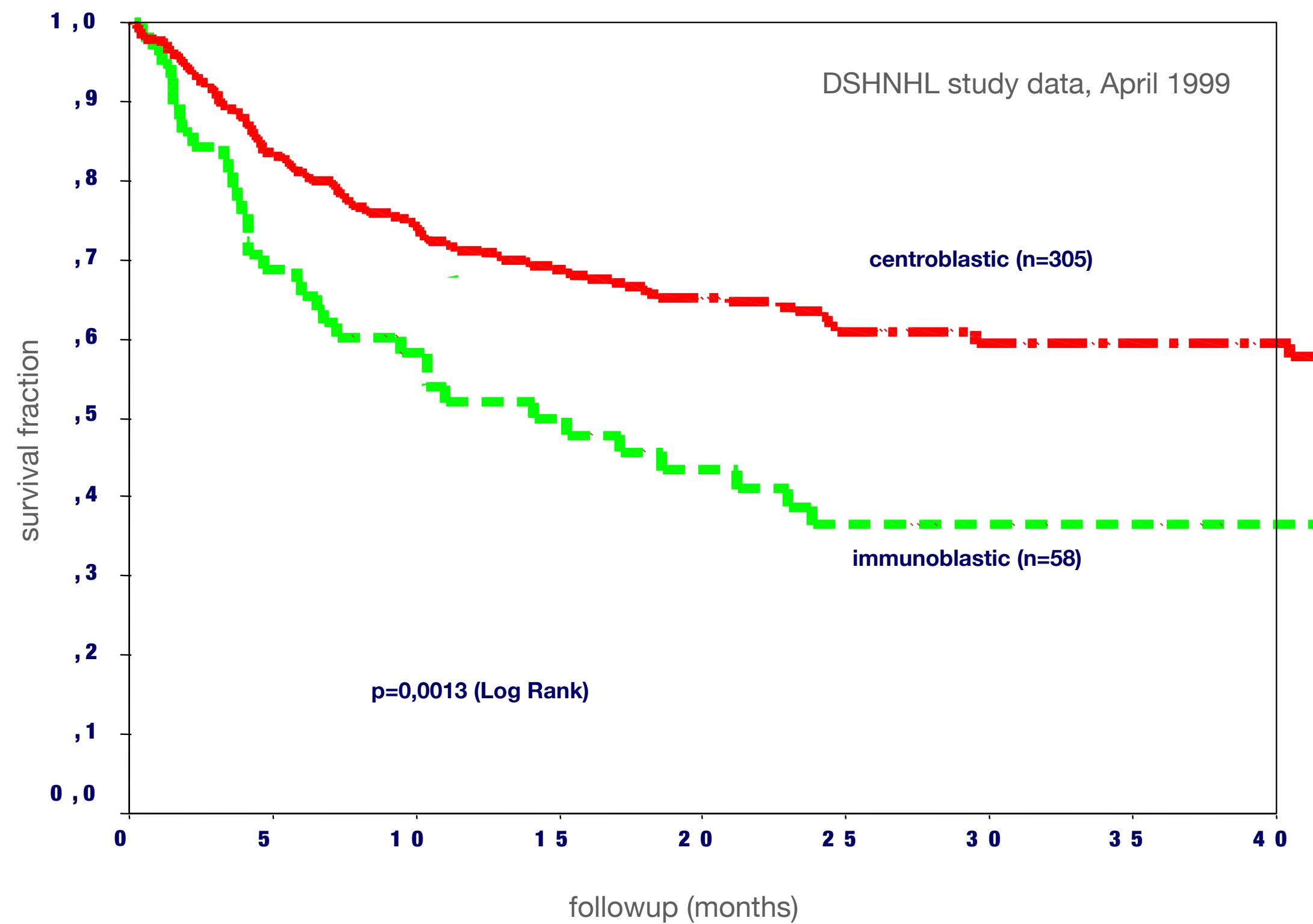
- ▶ Steps caused by multiplying by  $(1 - 1/49)$  and  $(1 - 1/46)$  respectively
- ▶ Late entry can also be dealt with

Kaplan-Meier estimators (km-na)

Kaplan-Meier estimators (km-na)

# Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets



# Kaplan-Meier Plots to Visualize Differential Risk

## Multi-parametric "risk scores" in CLL Prognosis

Leukemia (2020) 34:1038–1051  
<https://doi.org/10.1038/s41375-020-0727-y>

### ARTICLE

Chronic lymphocytic leukemia

#### Prognostic model for newly diagnosed CLL patients in Binet stage A: results of the multicenter, prospective CLL1 trial of the German CLL study group

Manuela A. Hockstetter<sup>1</sup> · Raymonde Busch<sup>2</sup> · Barbara Eichhorst<sup>3</sup> · Andreas Bühlert<sup>4</sup> · Dirk Winkler<sup>4</sup> · Jasmin Bahlo<sup>3</sup> · Sandra Robrecht<sup>3</sup> · Michael J. Eckart<sup>2</sup> · Ursula Vehling-Kaiser<sup>5</sup> · Georg Jacobs<sup>2</sup> · Ulrich Jäger<sup>8</sup> · Hans-Jürgen Hurtz<sup>2</sup> · Georg Hopfinger<sup>10</sup> · Frank Hartmann<sup>11</sup> · Harald Fuss<sup>12</sup> · Wolfgang Abenhards<sup>13</sup> · Ilona Blau<sup>14</sup> · Werner Freier<sup>15</sup> · Lothar Müller<sup>16</sup> · Maria Goebeler<sup>17</sup> · Clemens Wendtner<sup>1,3</sup> · Kirsten Fischer<sup>3</sup> · Carmen D. Herling<sup>3</sup> · Michael Stärk<sup>1</sup> · Martin Bentz<sup>18</sup> · Bertold Emmerich<sup>19</sup> · Hartmut Döhner<sup>20</sup> · Stephan Stilgenbauer<sup>20</sup> · Michael Hallek<sup>3</sup>

**Table 2a** Results of the Cox's regression for OS and TTFT in CLL patients in whom all 30 baseline parameters were available.

Univariate comparison	Hazard ratio [HR]	95% Confidence Interval		<i>P</i> value
		Lower	Upper	
<b>COX regression OS</b>				
Cytogenetic Hierarchical Type				
del(17p)	vs. not del(17p)/del(11q)	3.8	2.1	7.1
del(11q)	vs. not del(17p)/del(11q)	2.0	1.2	3.5
LDT				<0.001
<12 months	vs. ≥12 months	1.9	1.3	2.8
Age, years				0.001
>60	vs. ≤60	1.8	1.2	2.7
B2M, mg/dL				0.002
>3.5	vs. ≤3.5	2.0	1.2	3.1
IGHV mutational status				0.004
Unmutated	vs. mutated	2.4	1.6	3.6
<b>COX regression TTFT</b>				
Cytogenetic Hierarchical Type				
del(17p)	vs. not del(17p)/del(11q)	2.2	1.2	4.1
del(11q)	vs. not del(17p)/del(11q)	2.0	1.3	3.0
LDT	vs.	2.3	1.7	3.1
Age, years				<0.001
>60	vs. ≤60	1.3	1.0	1.7
B2M, mg/dL				0.037
>3.5	vs. ≤3.5	1.5	1.0	2.3
IGHV mutational status				0.049
Unmutated	vs. mutated	4.4	3.2	5.9
				<0.001

**Table 2b** Allocation of risk score points to the distinctive factors of the CLL1-PM.

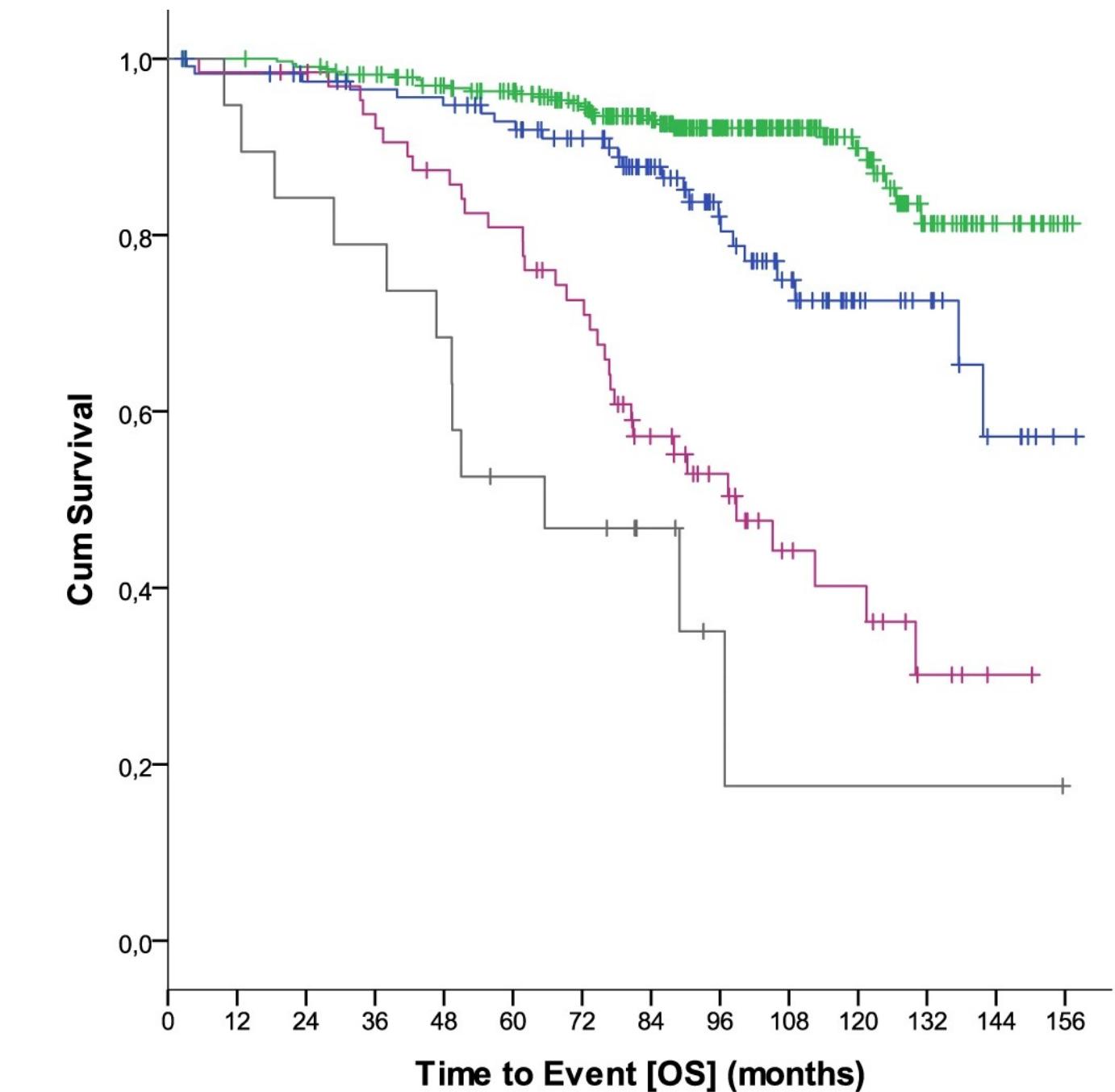
Characteristics	HR (95% CI)	<i>P</i>	Allocated risk score points
Del(17p)	3.8 (2.1–7.1)	<0.001	3.5
Unmutated <i>IGHV</i>	2.4 (1.6–3.6)	<0.001	2.5
Del(11q)	2.0 (1.2–3.5)	0.008	2.5
Beta2-MG >3.5 mg/L	2.0 (1.2–3.1)	0.004	2.5
LDT<12 months	1.9 (1.3–2.8)	0.001	1.5
Age >60 years	1.8 (1.2–2.7)	0.002	1.5

The assigned risk score points derived from the HR for OS of the individual factors.

**Table 2c** Patients and risk groups according to the CLL1 Prognostic Model (CLL1-PM). Patients and risk groups according to the CLL-IPI.

	Index score	Patients N (%)
Risk Groups according to the CLL1-PM		
Very low	0.0–1.5	336 (62.3)
Low	2.0–4.0	119 (22.1)
High	4.5–6.5	65 (12.1)
Very high	7.0–14.0	19 (3.5)
Risk Groups according to the CLL-IPI		
Low	0–1	360 (66.8)
Intermediate	2–3	141 (26.2)
High	4–6	33 (6.1)
Very high	7–10	5 (0.9)

OS overall survival, HR hazard ratio, Beta2-MG beta-2 microglobulin, *IGHV* immunoglobulin heavy-chain genes, LDT lymphocyte doubling time, TTFT time-to-first treatment.



***P* < 0.001**

Number at risk	0	12	24	36	48	60	72	84	96	108	120	132	144	156
<b>Very low</b>	336	335	331	322	306	294	262	215	160	113	68	33	15	2
<b>Low</b>	119	115	111	108	106	100	89	71	49	34	19	14	6	1
<b>High</b>	65	64	63	59	54	50	43	29	21	12	10	4	1	0
<b>Very high</b>	19	18	16	15	13	9	8	5	2	1	1	1	1	0

**Discrimination:** C-statistics, C = 0.739 (95% CI, 0.686–0.790)  
**AIC=445**

Overall survival according to the CLL1-PM risk groups. The full analysis dataset is comprised of the dataset of 539 patients.

# Kaplan-Meier Plots to Visualize Differential Risk

## Multi-parametric "risk scores" in CLL Prognosis

Leukemia (2020) 34:1038–1051  
<https://doi.org/10.1038/s41375-020-0727-y>

### ARTICLE

Chronic lymphocytic leukemia

#### Prognostic model for newly diagnosed CLL patients in Binet stage A: results of the multicenter, prospective CLL1 trial of the German CLL study group

Manuela A. Hockstetter<sup>1</sup> · Raymonde Busch<sup>2</sup> · Barbara Eichhorst<sup>3</sup> · Andreas Bühlert<sup>4</sup> · Dirk Winkler<sup>4</sup> · Jasmin Bahlo<sup>3</sup> · Sandra Robrecht<sup>3</sup> · Michael J. Eckart<sup>5</sup> · Ursula Vehling-Kaiser<sup>6</sup> · Georg Jacobs<sup>7</sup> · Ulrich Jäger<sup>8</sup> · Hans-Jürgen Hurtz<sup>2</sup> · Georg Hopfinger<sup>10</sup> · Frank Hartmann<sup>11</sup> · Harald Fuss<sup>12</sup> · Wolfgang Abenhards<sup>13</sup> · Ilona Blau<sup>14</sup> · Werner Freier<sup>15</sup> · Lothar Müller<sup>16</sup> · Maria Goebeler<sup>17</sup> · Clemens Wendtner<sup>1,3</sup> · Kirsten Fischer<sup>3</sup> · Carmen D. Herling<sup>3</sup> · Michael Stärck<sup>1</sup> · Martin Bentz<sup>18</sup> · Bertold Emmerich<sup>19</sup> · Hartmut Döhner<sup>20</sup> · Stephan Stilgenbauer<sup>20</sup> · Michael Hallek<sup>3</sup>

**Table 2a** Results of the Cox's regression for OS and TTFT in CLL patients in whom all 30 baseline parameters were available.

Univariate comparison	Hazard ratio [HR]	95% Confidence Interval		<i>P</i> value
		Lower	Upper	
<b>COX regression OS</b>				
Cytogenetic Hierarchical Type				
del(17p) vs. not del(17p)/del(11q)	3.8	2.1	7.1	<0.001
del(11q) vs. not del(17p)/del(11q)	2.0	1.2	3.5	0.008
LDT				
<12 months vs. ≥12 months	1.9	1.3	2.8	0.001
Age, years				
>60 vs. ≤60	1.8	1.2	2.7	0.002
B2M, mg/dL				
>3.5 vs. ≤3.5	2.0	1.2	3.1	0.004
IGHV mutational status				
Unmutated vs. mutated	2.4	1.6	3.6	<0.001
<b>COX regression TTFT</b>				
Cytogenetic Hierarchical Type				
del(17p) vs. not del(17p)/del(11q)	2.2	1.2	4.1	0.009
del(11q) vs. not del(17p)/del(11q)	2.0	1.3	3.0	0.001
LDT				
vs. <12 months	2.3	1.7	3.1	<0.001
Age, years				
>60 vs. ≤60	1.3	1.0	1.7	0.037
B2M, mg/dL				
>3.5 vs. ≤3.5	1.5	1.0	2.3	0.049
IGHV mutational status				
Unmutated vs. mutated	4.4	3.2	5.9	<0.001

**Table 2b** Allocation of risk score points to the distinctive factors of the CLL1-PM.

Characteristic	HR (95% CI)	<i>P</i>	Allocated risk score points
Del(17p)	3.8 (2.1–7.1)	<0.001	3.5
Del(11q)	2.0 (1.2–3.5)	0.008	2.5
Beta2-MG >3.5 mg/L	2.0 (1.2–3.1)	0.004	2.5
LDT<12 months	1.9 (1.3–2.8)	0.001	1.5
Age >60 years	1.8 (1.2–2.7)	0.002	1.5

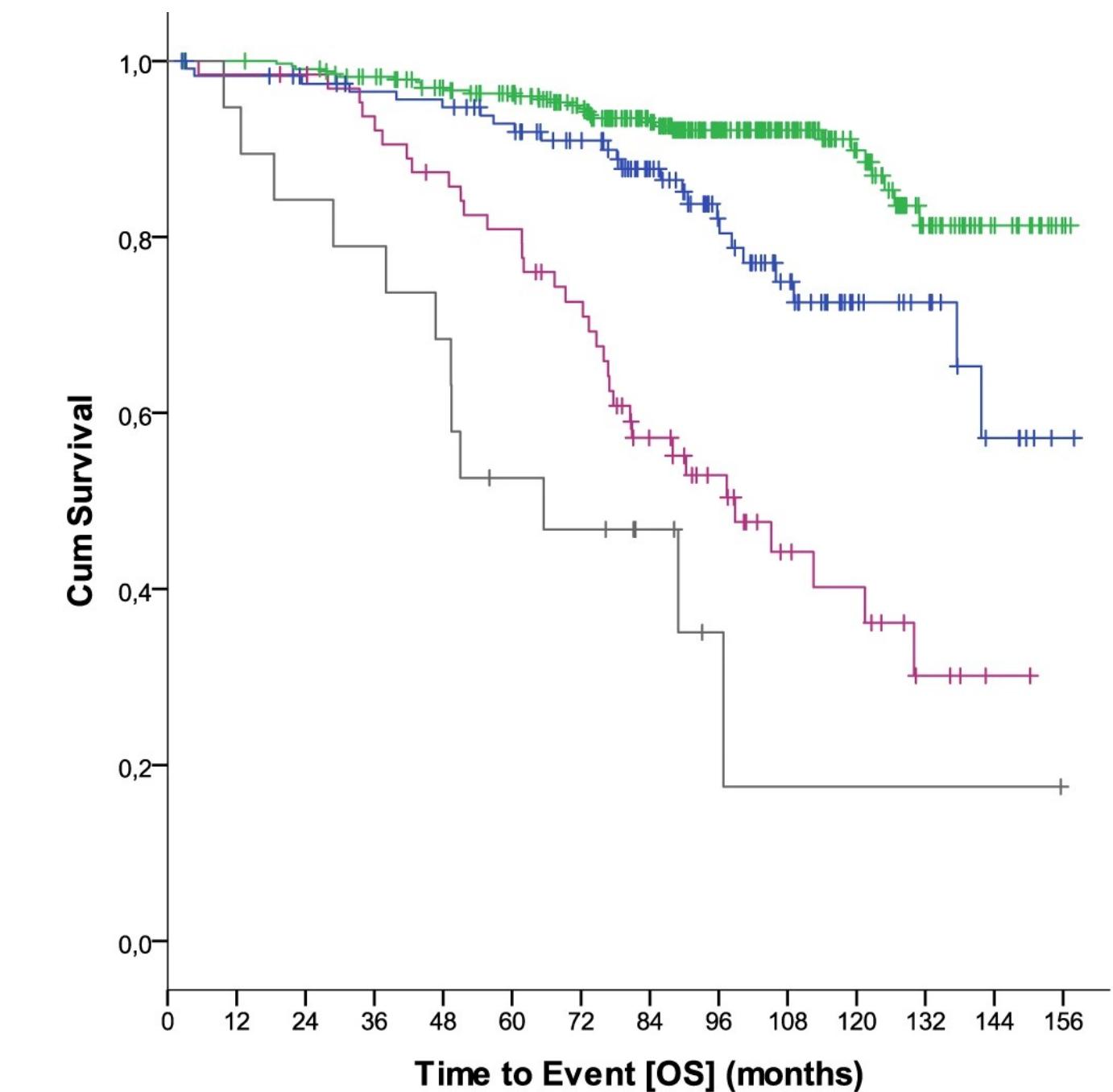
The assigned risk score points derived from the HR for OS of the individual factors.

**Table 2c** Patients and risk groups according to the CLL1 Prognostic Model (CLL1-PM). Patients and risk groups according to the CLL-IPI.

	Index score	Patients N (%)
Risk Groups according to the CLL1-PM		
Very low	0.0–1.5	336 (62.3)
Low	2.0–4.0	119 (22.1)
High	4.5–6.5	65 (12.1)
Very high	7.0–14.0	19 (3.5)
Risk Groups according to the CLL-IPI		
Low	0–1	360 (66.8)
Intermediate	2–3	141 (26.2)
High	4–6	33 (6.1)
Very high	7–10	5 (0.9)

OS overall survival, HR hazard ratio, Beta2-MG beta-2 microglobulin, IGHV immunoglobulin heavy-chain genes, LDT lymphocyte doubling time, TTFT time-to-first treatment.

- "a novel prognostic model (CLL1-PM) developed to identify risk groups, separating patients with favorable from others with dismal prognosis"
- " findings would be useful to effectively stratify Binet stage A patients, particularly within the scope of clinical trials evaluating novel agents"



***P* < 0.001**

Number at risk	0	12	24	36	48	60	72	84	96	108	120	132	144	156
<b>Very low</b>	336	335	331	322	306	294	262	215	160	113	68	33	15	2
<b>Low</b>	119	115	111	108	106	100	89	71	49	34	19	14	6	1
<b>High</b>	65	64	63	59	54	50	43	29	21	12	10	4	1	0
<b>Very high</b>	19	18	16	15	13	9	8	5	2	1	1	1	1	0

**Discrimination:** C-statistics, C = 0.739 (95% CI, 0.686–0.790)  
**AIC=445**

Overall survival according to the CLL1-PM risk groups. The full analysis dataset is comprised of the dataset of 539 patients.

# Cancer Classifications & Parameters

NCIt | ICD-O / WHO | TNM

# ICD-O 3

## WHO International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3)

- used in cancer registries for coding the site (topography) and the histology (morphology) of neoplasms, usually obtained from a pathology report
- mix of "biology" (i.e. tumor morphology) and "clinical" (i.e. tumor site)

→ 2 codes per cancer

▶ "Adenocarcinoma" of the "Sigmoid colon"

**8140/3**

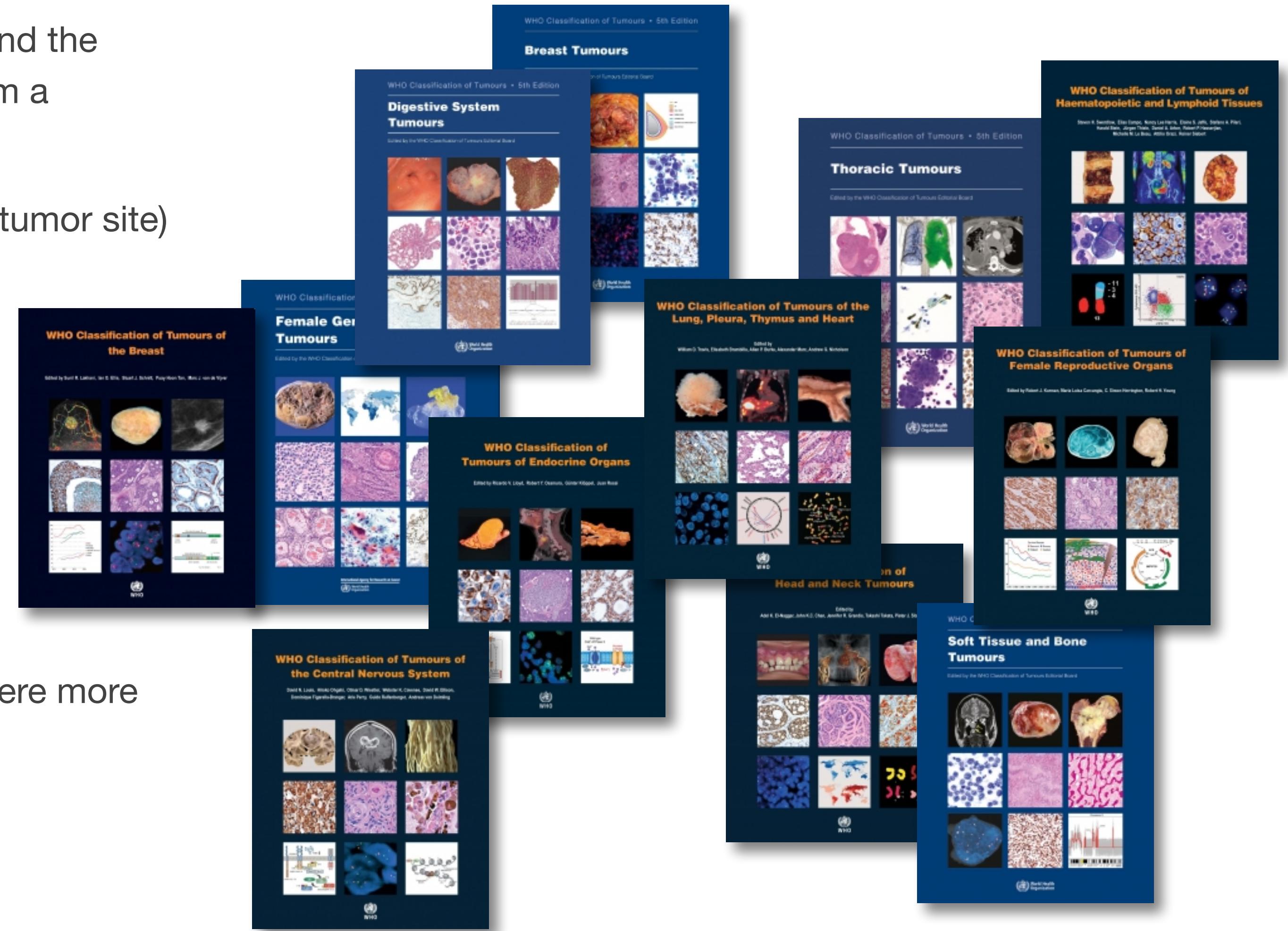
**C18.7**

▶ "Retinoblastoma" of the "Retina"

**9510/3**

**C69.2**

- widely accepted by pathologists but limited clinical use (there more ICD-10 or SNOMED)
- no ontology & not (truly) hierarchical
- many entities difficult to remap if using only single code



# NCIt

## Neoplasm Classifications in the NCI Thesaurus

- NCI's core reference terminology and biomedical ontology are collected in the NCI Thesaurus (NCIt)
- individual codes for site-specific occurrences of "biological" diagnoses

### 1 code per cancer

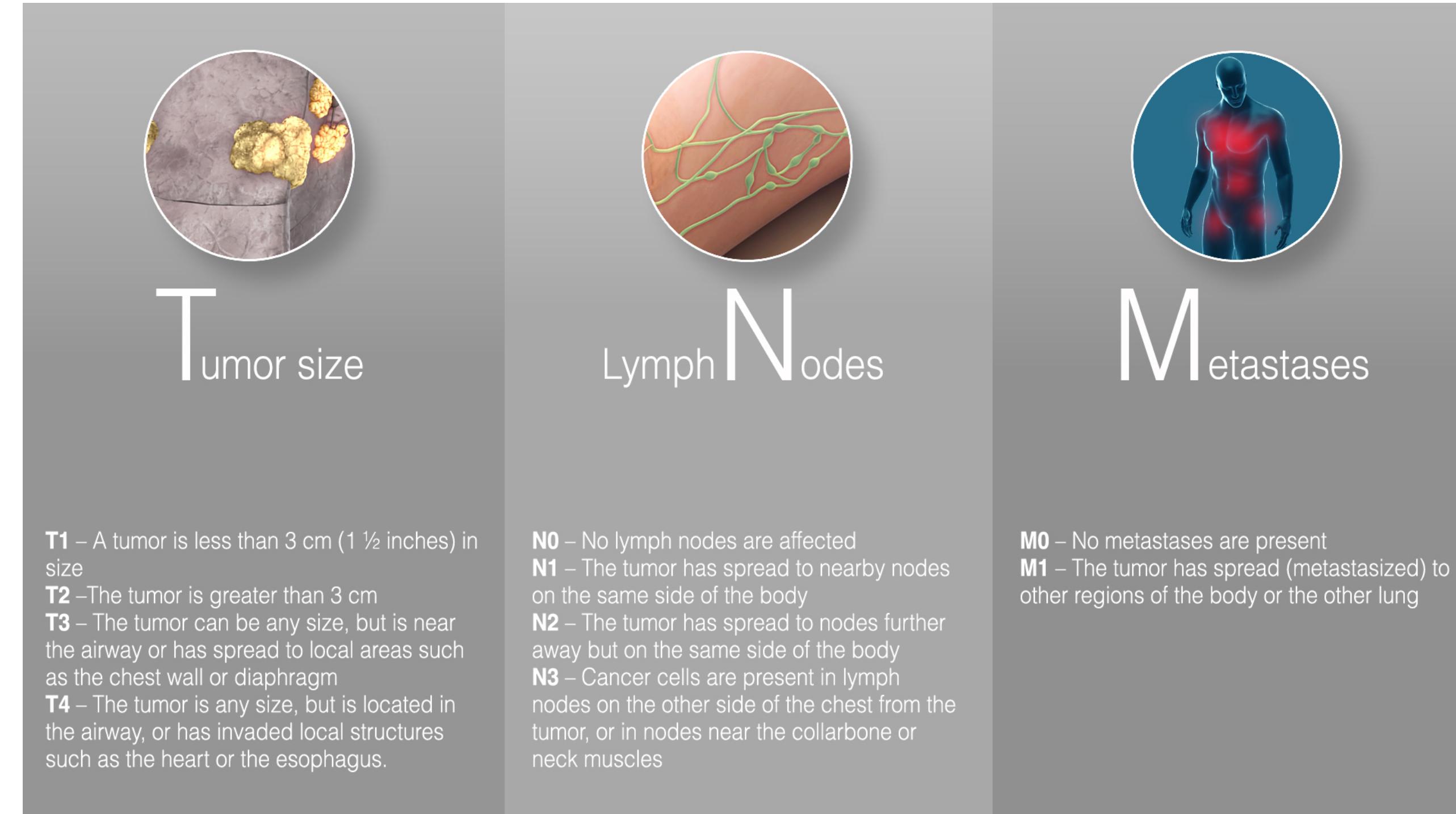
- ▶ **NCIT:C43584** - Rectosigmoid Adenocarcinoma
- ▶ **NCIT:C7541** - Retinoblastoma
- truly hierarchical ontology
- hierarchical system empowers "logical OR" queries
- terms can have multiple occurrences in diagnostic tree
- assignment of code to different groupings allows soft aggregation (e.g. a type of colorectal adenocarcinoma with all colon tumors or with all adenocarcinomas)

- ▼ NCIT:C3262: Neoplasm (116013 samples)
- ▼ NCIT:C3263: Neoplasm by Site (110893 samples)
  - NCIT:C156482: Genitourinary System Neoplasm (16534 samples)
  - NCIT:C2910: Breast Neoplasm (15957 samples)
  - NCIT:C3010: Endocrine Neoplasm (3521 samples)
  - NCIT:C3030: Eye Neoplasm (280 samples)
- ▼ NCIT:C3052: Digestive System Neoplasm (15289 samples)
  - NCIT:C172852: Digestive System Soft Tissue Neoplasm (99 samples)
  - NCIT:C27721: Digestive System Neuroendocrine Neoplasm (202 samples)
  - NCIT:C2877: Anal Neoplasm (61 samples)
  - NCIT:C3028: Esophageal Neoplasm (1865 samples)
- ▼ NCIT:C3141: Intestinal Neoplasm (5723 samples)
  - ▼ NCIT:C2956: Colorectal Neoplasm (5579 samples)
    - NCIT:C2953: Colon Neoplasm (4666 samples)
    - NCIT:C3350: Rectal Neoplasm (527 samples)
    - NCIT:C4610: Benign Colorectal Neoplasm (181 samples)
  - ▼ NCIT:C4877: Rectosigmoid Neoplasm (240 samples)
    - ▼ NCIT:C7420: Malignant Rectosigmoid Neoplasm (240 samples)
    - ▼ NCIT:C7421: Rectosigmoid Carcinoma (240 samples)
  - ▼ NCIT:C43584: Rectosigmoid Adenocarcinoma (240 samples)
    - NCIT:C43592: Rectosigmoid Mucinous Adeno... (18 samples)
  - NCIT:C4978: Malignant Colorectal Neoplasm (5398 samples)
  - NCIT:C96152: Colorectal Neuroendocrine Neoplasm (11 samples)
  - NCIT:C4432: Small Intestinal Neoplasm (66 samples)

# TNM

## A Classification for Clinical Cancer Stage Parameters

- most widely used cancer staging system
- **T** refers to the size and extent of the main tumor
- **N** refers to the the number / location of nearby lymph nodes that have cancer infiltration
- **M** refers to whether the cancer has metastasized
- not used for leukemias / lymphomas
  - Binet and Rai in CLL
  - proportion of blasts in bone marrow or blood in leukemias
  - Lugano classification in lymphomas
- other disease specific staging systems may (co-) exist
  - e.g. a stage II breast cancer is determined by size & nodal involvement



Source: [www.scientificanimations.com](http://www.scientificanimations.com)

# TNM

## A Classification for Clinical Cancer Stage Parameters

- most widely used cancer staging system
- T** refers to the size and extent of the main tumor
- N** refers to the the number / location of nearby lymph nodes that have cancer infiltration
- M** refers to whether the cancer has metastasized
- not used for leukemias / lymphomas
  - Binet and Rai in CLL
  - proportion of blasts in bone marrow or blood in leukemias
  - Lugano classification in lymphomas
- other disease specific staging systems may (co-) exist
  - e.g. a stage II lung cancer is determined by size & nodal involvement

## TNM STAGING OF LUNG CANCER - 8<sup>th</sup> EDITION

The diagram illustrates the TNM Staging of Lung Cancer - 8<sup>th</sup> Edition. It starts with a legend for M status:

- Stage IV B (Any T, Any N, M1c)**: Multiple extrathoracic metastases (in one or more organs)
- Stage IV A (Any T, Any N, M1a/b)**: Single extrathoracic metastasis (including non-regional lymph nodes)
- M1b**: Satellite (separate) tumor nodule(s) in contralateral lobe or Pleural or pericardial nodules or malignant effusion
- M1a**
- Distant metastasis M1**
- No distant metastasis M0**

Arrows point from M1c and M1a/b to Stage IV B and Stage IV A respectively, and from M1 and M0 to Distant metastasis and No distant metastasis.

**LYMPH NODE (N)**

Scalene (ipsi/ contralateral)	Supravacular	Hilar	Medastinal	Subcarinal	Medastinal	Hilar	Ipsilateral	Peribronchial
● ● ● ●		- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -

**Explanation of lymph node staging:**

- For any N category, one or more of the groups marked by ● must be involved and the involvement of all groups marked by - should be absent.
- The presence or absence of involvement in groups marked by □ does not alter N staging in the corresponding category.

**PRIMARY TUMOR (T)**

Stage I A1 (T1 (mi) N0 M0)		Stage 0 (Tis N0 M0)		Occult Carcinoma (Tx N0 M0)		1- Size (greatest dimension)		2- Criteria of Extent		Endo-bronchial Location		Local Invasion		Separate Tumor Nodule(s)			
T1a	T1b	T1c	T2a	T2b	T1	T2	T1a	T1b	T1c	T2a	T2b	T1	T2	Absent	Absent		
T1a	T1b	T1c	T2a	T2b	T1	T2	≤ 1 cm	> 1 cm ≤ 2 cm	> 2 cm ≤ 3 cm	> 3 cm ≤ 4 cm	> 4 cm ≤ 5 cm	or Any size ≤ 5 cm in the presence of 1 or more of the criteria of extent*	> 5 cm ≤ 7 cm	or Any size ≤ 7 cm in the presence of 1 or more of the criteria of extent	> 7 cm		

**Legend for Stage Categories:**

- Stage I A1 (T1 (mi) N0 M0)**: T1 (mi): Minimally invasive adenocarcinoma (solitary adenocarcinoma, ≤ 3 cm with a lepidic growth and ≤ 5 mm invasion in any focus)
- Stage 0 (Tis N0 M0)**: Tis: Carcinoma in situ
- Occult Carcinoma (Tx N0 M0)**: Tx: Tumor is proven histopathologically (+ Cytology) but not detected by imaging or bronchoscopy

Lababede O, Meziane MA. The Eighth Edition of TNM Staging of Lung Cancer: Reference Chart and Diagrams. Oncologist. 2018;23(7):844-848.

# TNM

## A Classification for Clinical Cancer Stage Parameters

- most widely used cancer staging system
- **T** refers to the size and extent of the main tumor
- **N** refers to the the number / location of nearby lymph nodes that have cancer infiltration
- **M** refers to whether the cancer has metastasized
- not used for leukemias / lymphomas
  - ▶ Binet and Rai in CLL
  - ▶ proportion of blasts in bone marrow or blood in leukemias
  - ▶ Lugano classification in lymphomas
- other disease specific staging systems may (co-) exist
  - ▶ e.g. a stage II lung cancer is determined by size & nodal involvement

TNM has been  
"ontologized"  
into NCIt

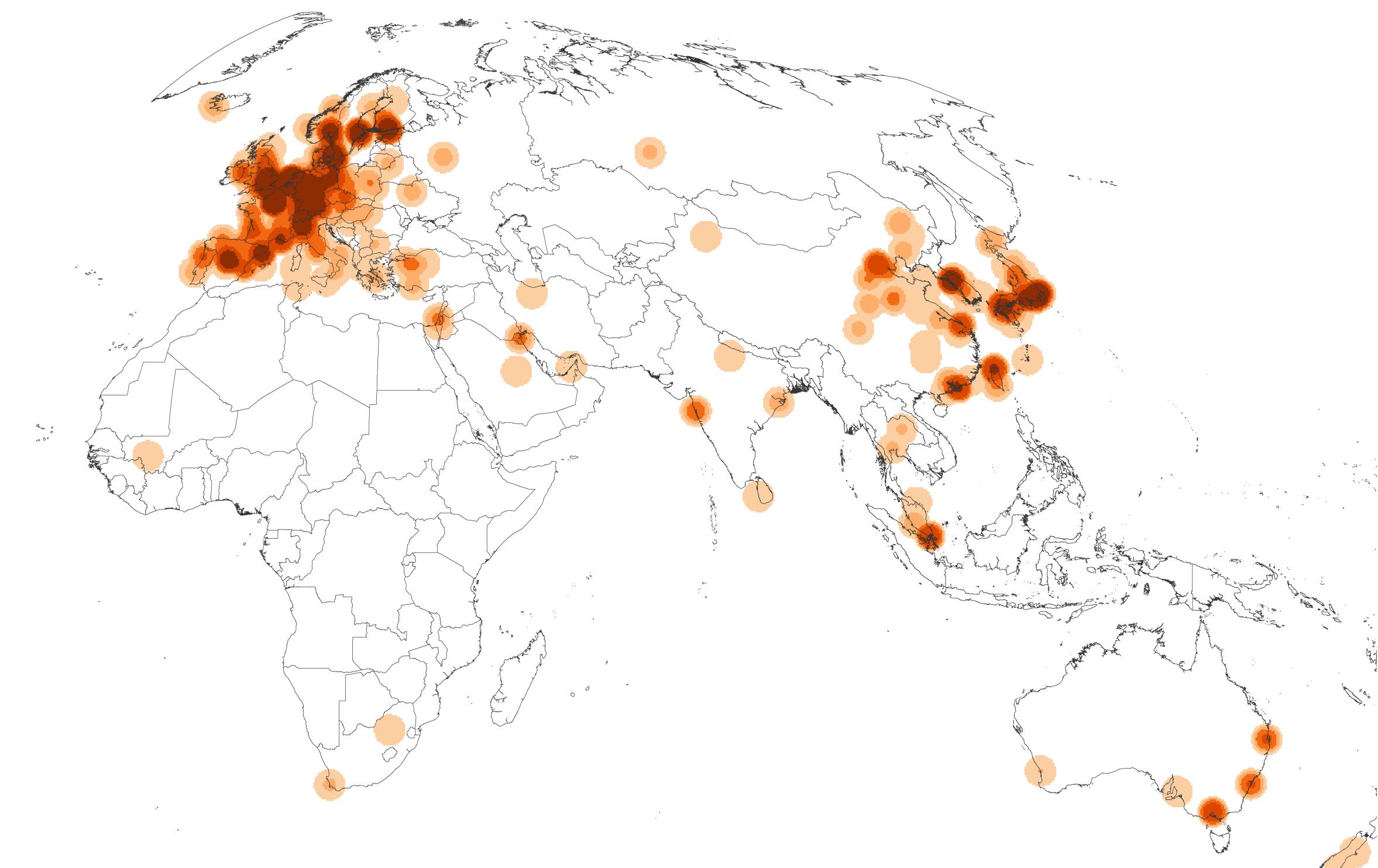
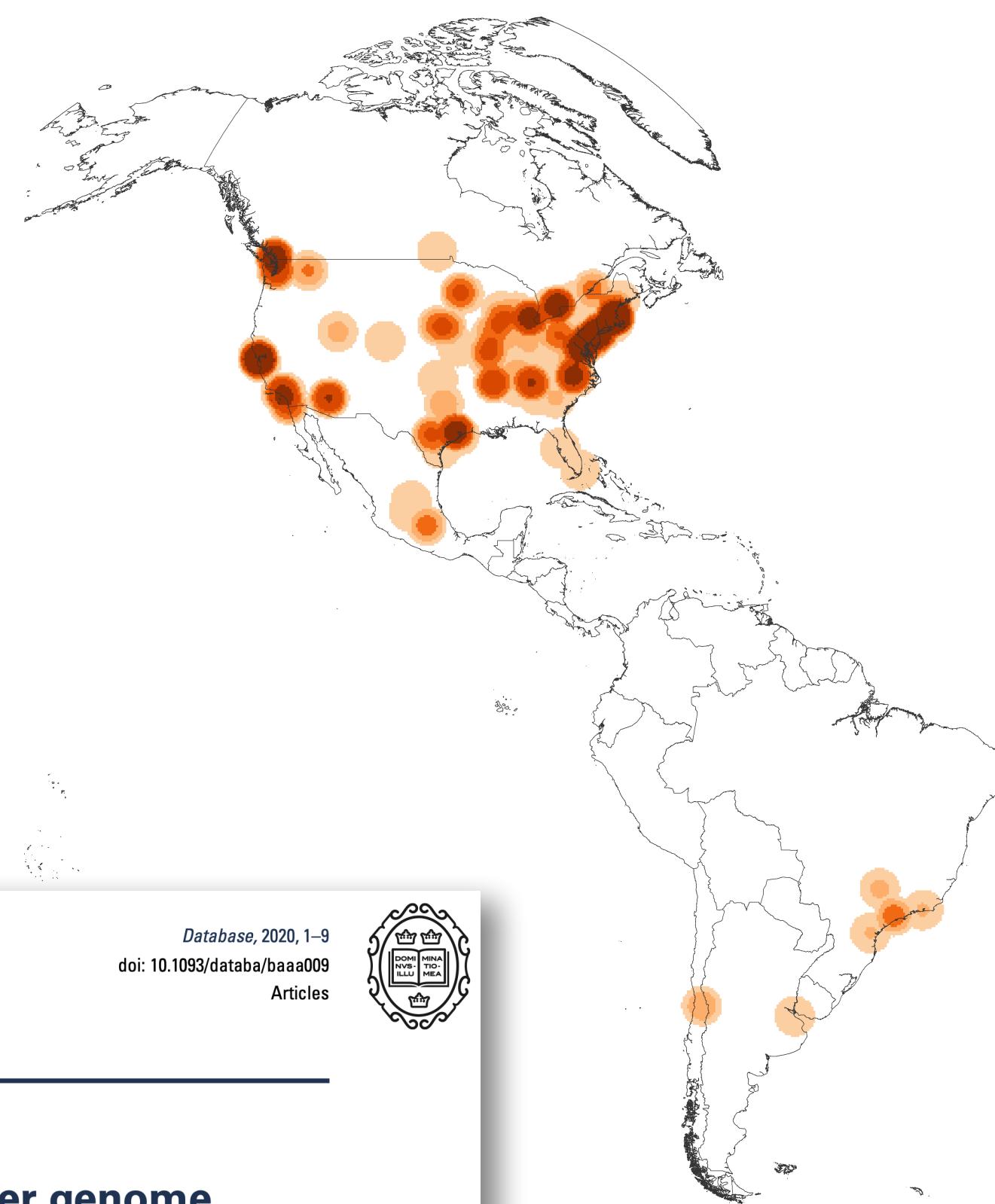
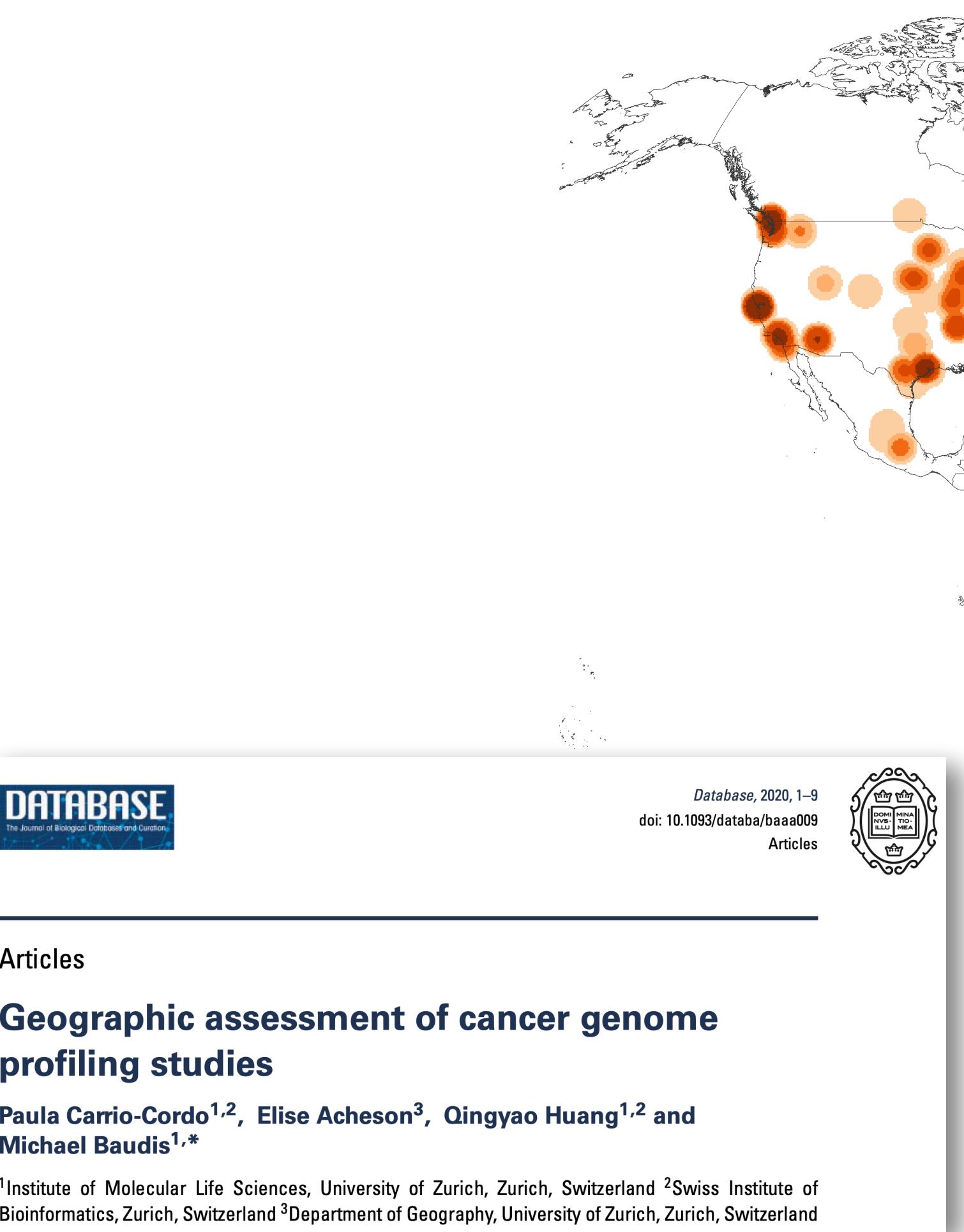
NCIT:C48698	Cancer TNM Finding Category	0
NCIT:C133398	Postneoadjuvant Therapy Pathologic TNM Finding	1
NCIT:C143081	Posttherapy Clinical TNM Finding	1
NCIT:C48739	Pathologic TNM Finding	1
NCIT:C48886	Pathologic Distant Metastasis TNM Finding	2
NCIT:C48740	pM0 Stage Finding	3
NCIT:C48741	pM1 Stage Finding	3
NCIT:C48742	pM1a Stage Finding	4
NCIT:C48743	pM1b Stage Finding	4
NCIT:C48744	pM1c Stage Finding	4
NCIT:C48887	Pathologic Regional Lymph Nodes TNM Finding	2
NCIT:C48745	pN0 Stage Finding	3
NCIT:C48746	pN1 Stage Finding	3
NCIT:C48747	pN1a Stage Finding	4
NCIT:C48748	pN1b Stage Finding	4
NCIT:C48749	pN1c Stage Finding	4
NCIT:C48750	pN2 Stage Finding	3
NCIT:C48751	pN2a Stage Finding	4
NCIT:C48752	pN2b Stage Finding	4
NCIT:C48753	pN2c Stage Finding	4
NCIT:C48754	pN3 Stage Finding	3
NCIT:C48755	pN3a Stage Finding	4
NCIT:C48756	pN3b Stage Finding	4
NCIT:C48757	pN3c Stage Finding	4
NCIT:C48888	Pathologic Primary Tumor TNM Finding	2
NCIT:C48758	pT0 Stage Finding	3
NCIT:C48759	pT1 Stage Finding	3
NCIT:C48760	pT1a Stage Finding	4
NCIT:C48761	pT1b Stage Finding	4
NCIT:C48763	pT1c Stage Finding	4
NCIT:C48764	pT2 Stage Finding	3
NCIT:C48765	pT2a Stage Finding	4
NCIT:C48766	pT2b Stage Finding	4
NCIT:C48767	pT2c Stage Finding	4
NCIT:C48768	pT3 Stage Finding	3
NCIT:C48769	pT3a Stage Finding	4
NCIT:C48770	pT3b Stage Finding	4
NCIT:C48771	pT3c Stage Finding	4
NCIT:C48772	pT4 Stage Finding	3
NCIT:C48773	pT4a Stage Finding	4
NCIT:C48774	pT4b Stage Finding	4
NCIT:C48775	pT4c Stage Finding	4
NCIT:C48776	pT4d Stage Finding	4
NCIT:C48879	Generic TNM Finding	1
NCIT:C48777	Cancer TNM Vessel Invasion Finding Category	2
NCIT:C147091	Lymphovascular Invasion 0	3
NCIT:C147092	Lymphovascular Invasion 1	3
NCIT:C147093	Lymphovascular Invasion 9	3
NCIT:C147094	Lymphovascular Invasion 2	3
NCIT:C147095	Lymphovascular Invasion 3	3
NCIT:C147096	Lymphovascular Invasion 4	3
NCIT:C48883	Generic Distant Metastasis TNM Finding	2
NCIT:C48899	M0 Stage Finding	3
NCIT:C95956	cM0 (i+) Stage Finding	4
NCIT:C48700	M1 Stage Finding	3
NCIT:C48701	M1a Stage Finding	4
NCIT:C48702	M1b Stage Finding	4
NCIT:C48703	M1c Stage Finding	4
NCIT:C48704	MX Stage Finding	0
NCIT:C48884	Generic Regional Lymph Nodes TNM Finding	1
NCIT:C48705	N0 Stage Finding	1
NCIT:C95921	N0 (i-) Stage Finding	1
NCIT:C95922	N0 (i+) Stage Finding	2
NCIT:C95923	N0 (mol-) Stage Finding	3
NCIT:C95925	N0 (mol+) Stage Finding	3
NCIT:C48706	N1 Stage Finding	4
NCIT:C48707	N1a Stage Finding	4
NCIT:C48708	N1b Stage Finding	4
NCIT:C95929	N1bl Stage Finding	2
NCIT:C95935	N1bli Stage Finding	3
NCIT:C95936	N1bll Stage Finding	3
NCIT:C95937	N1bIV Stage Finding	4
NCIT:C48709	N1c Stage Finding	4
NCIT:C95955	N1mi Stage Finding	4
NCIT:C48714	N3 Stage Finding	3
NCIT:C48715	N3a Stage Finding	4
NCIT:C48716	N3b Stage Finding	4
NCIT:C48717	N3c Stage Finding	4
NCIT:C48718	NX Stage Finding	3
NCIT:C48786	N2 Stage Finding	4
NCIT:C48711	N2a Stage Finding	4
NCIT:C48712	N2b Stage Finding	4
NCIT:C48713	N2c Stage Finding	2
NCIT:C96026	N4 Stage Finding	3
NCIT:C48885	Generic Primary Tumor TNM Finding	3
NCIT:C106299	Any T	4
NCIT:C132010	T5 Stage Finding	4
NCIT:C48719	T0 Stage Finding	4
NCIT:C48720	T1 Stage Finding	3
NCIT:C48721	T1a Stage Finding	4
NCIT:C48722	T1b Stage Finding	4
NCIT:C48723	T1c Stage Finding	4
NCIT:C95805	T1mi Stage Finding	3
NCIT:C48724	T2 Stage Finding	4
NCIT:C148411	T2d Stage Finding	4
NCIT:C48725	T2a Stage Finding	4
NCIT:C48726	T2b Stage Finding	3
NCIT:C48727	T2c Stage Finding	4
NCIT:C48728	T3 Stage Finding	4
NCIT:C148412	T3d Stage Finding	4
NCIT:C48729	T3a Stage Finding	4
NCIT:C48730	T3b Stage Finding	1
NCIT:C48731	T3c Stage Finding	2
NCIT:C48732	T4 Stage Finding	3
NCIT:C48733	T4a Stage Finding	3
NCIT:C48734	T4b Stage Finding	3
NCIT:C48735	T4c Stage Finding	3
NCIT:C48736	T4d Stage Finding	3
NCIT:C48737	TX Stage Finding	3
NCIT:C48738	Tis Stage Finding	2
NCIT:C96025	Ta Stage Finding	3
NCIT:C48880	Recurrent Cancer TNM Finding	4
NCIT:C48881	Clinical TNM Finding	3
NCIT:C161009	Clinical Primary Tumor TNM Finding	4
NCIT:C162609	Clinical Regional Lymph Nodes TNM Finding	4
NCIT:C162610	Clinical Distant Metastasis TNM Finding	4
NCIT:C48882	Autopsy TNM Finding	4

# Tasks

## Survival analyses | Cancer classifications | Staging

- Familiarize yourself with the different concepts behind different disease classification systems - what are there use, advantages, problems? E.g. ICD-10, ICD-O, NCI  
  - you can use Progenetix to explore e.g. ontology mapping
- Learn to "read" Kaplan-Meier plots (preparation for explorative analyses later in the course).
- Achieve a principal understanding of TNM codes & write some "translations"  
→ T1N1M0: small tumor with regional lymph node involvement and no detected distant metastases

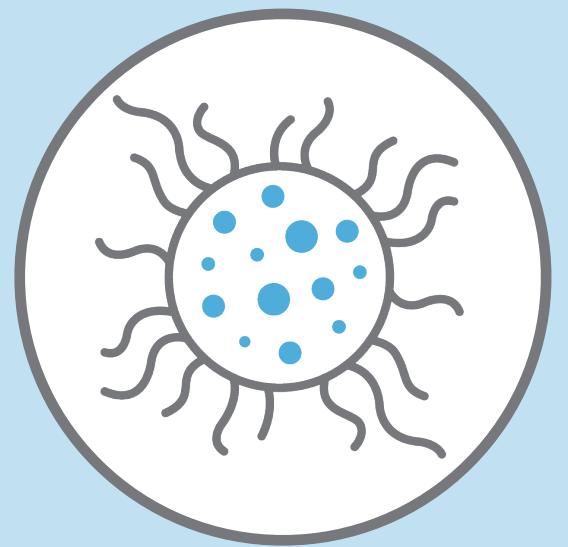
# Where does Genomic Data Come From?



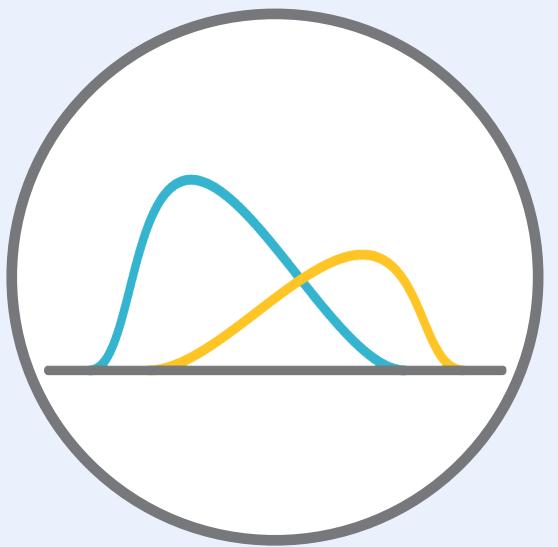
Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets. The numbers are derived from the 3'240 publications registered in the Progenetix database.



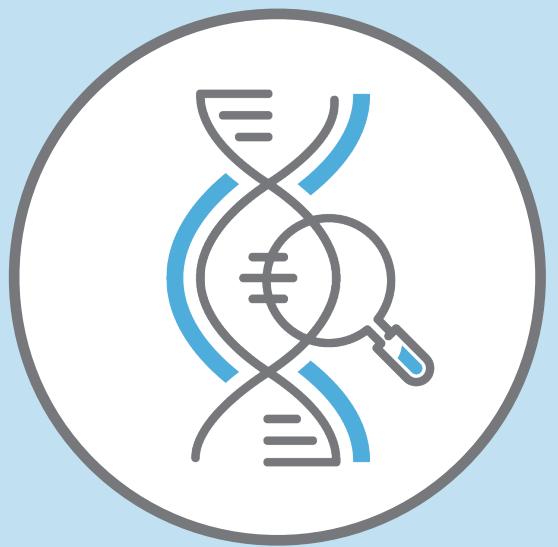
# Global Genomic Data Sharing Can...



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



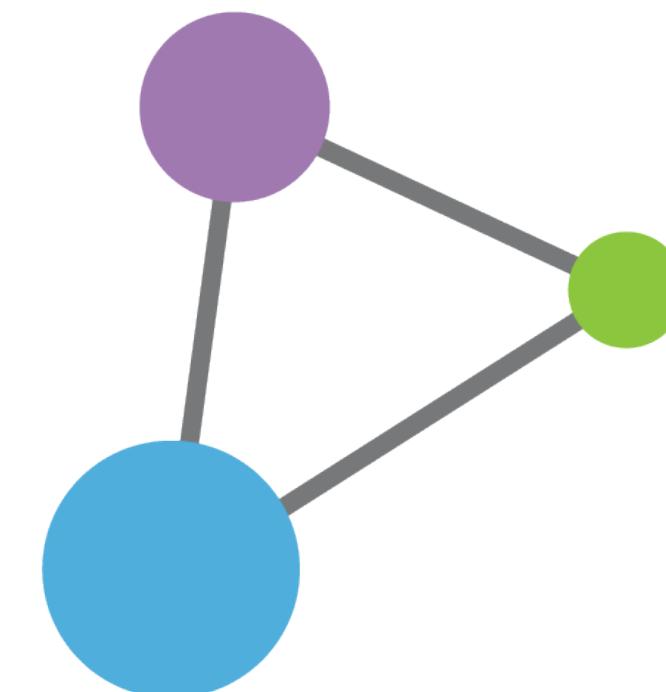
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

# Different Approaches to Data Sharing



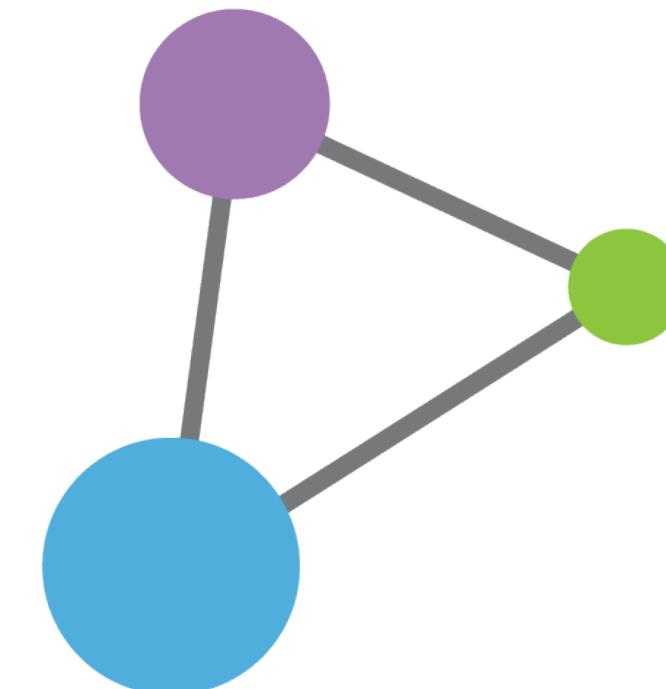
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets

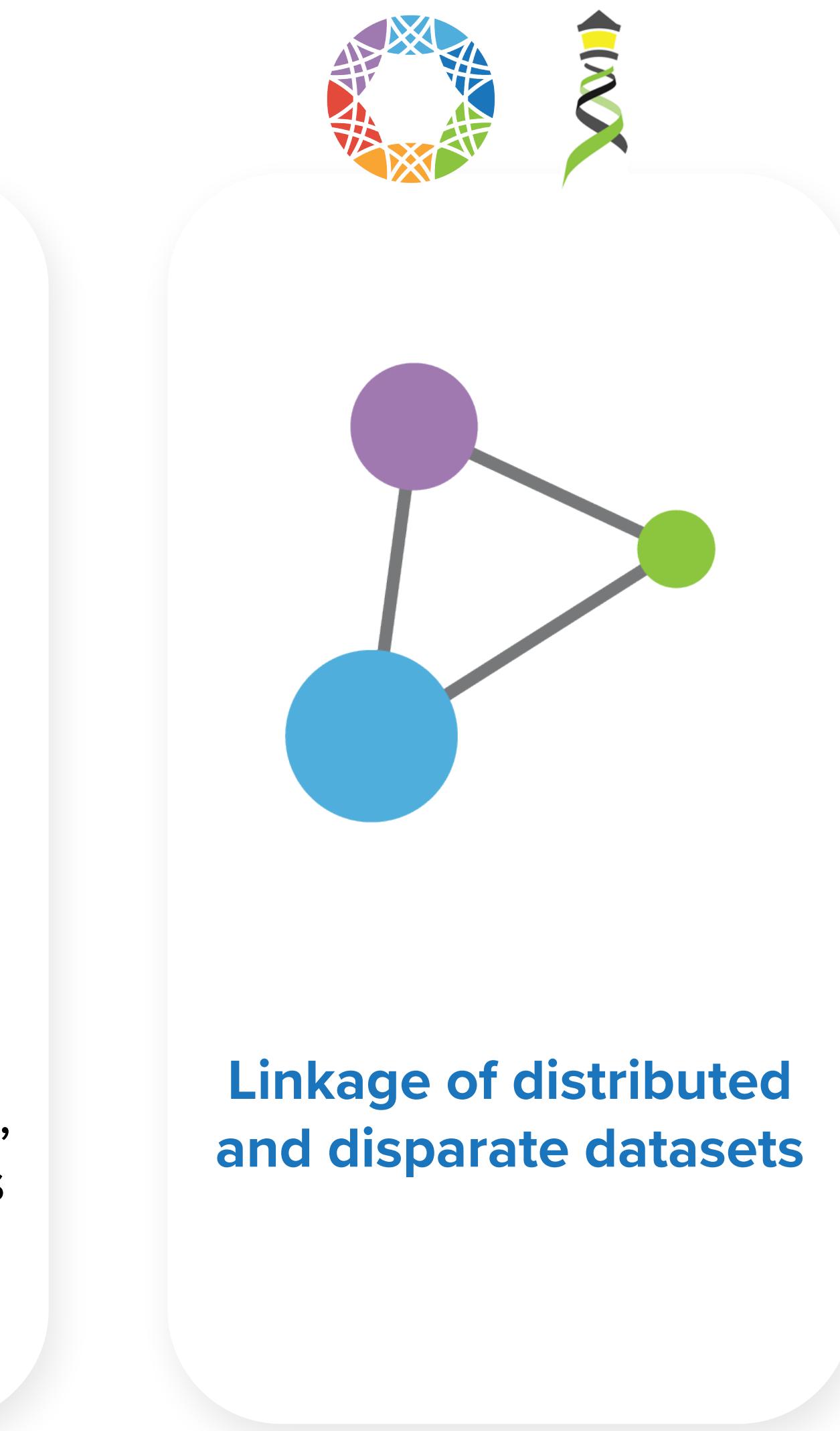
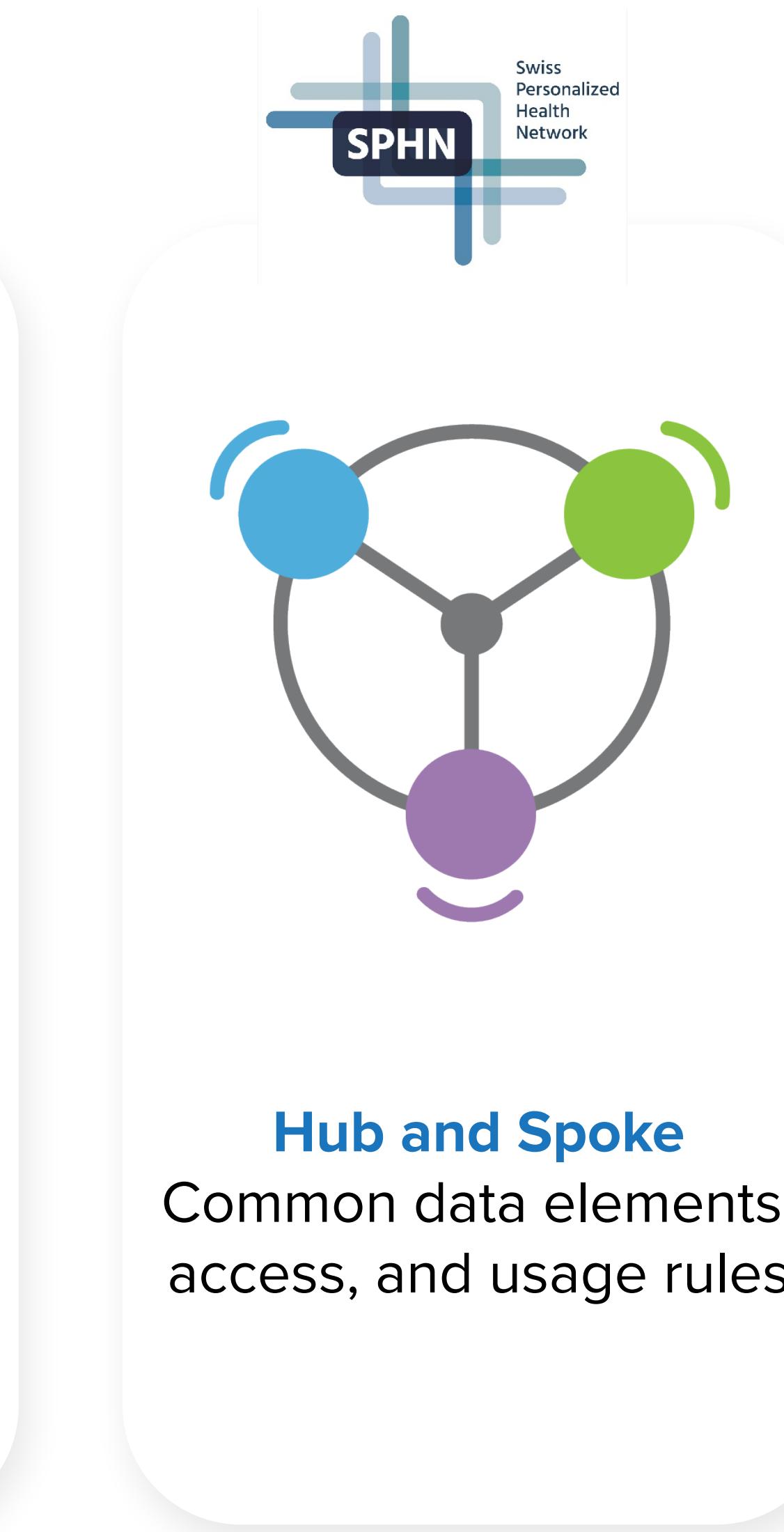
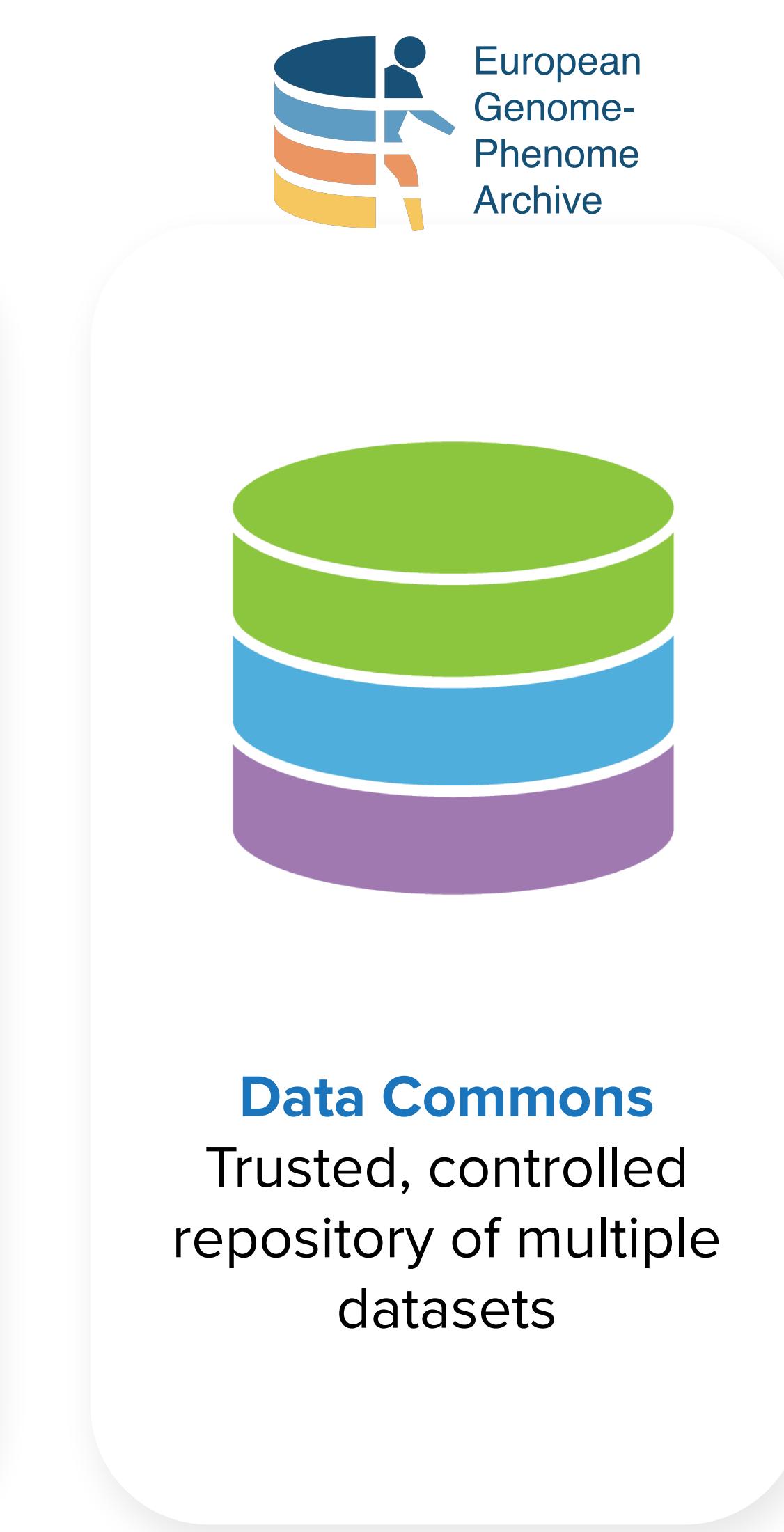


**Hub and Spoke**  
Common data elements, access, and usage rules



**Linkage of distributed and disparate datasets**

# Different Approaches to Data Sharing



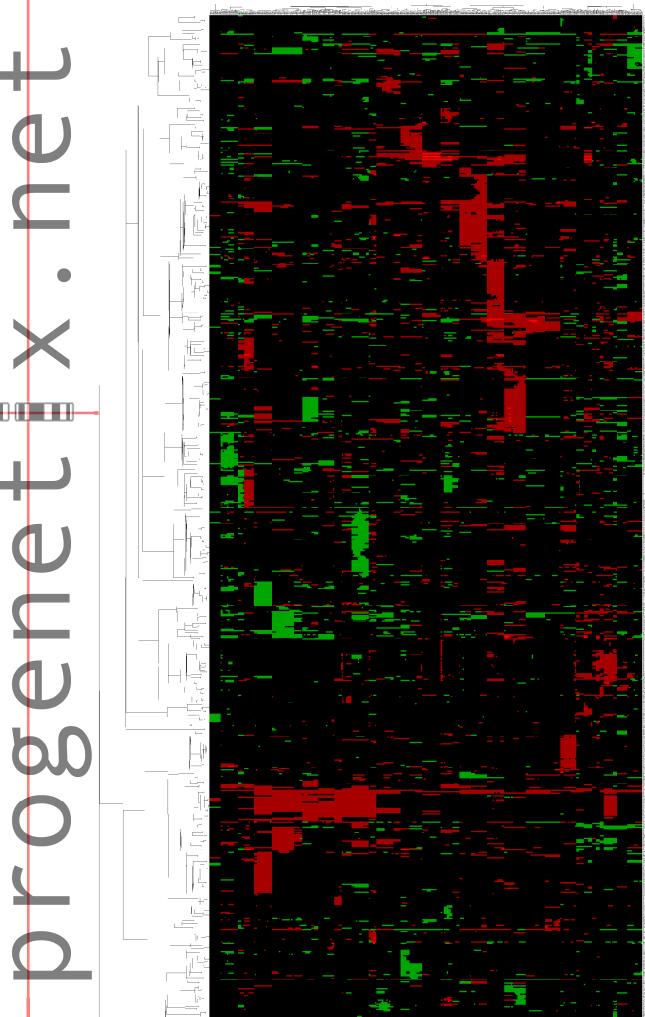
**progenetix.net: storage and visualization of genomic aberration data in human malignancies**  
michael baudis, md

Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic "hotspots". Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under [www.progenetix.net/bcats/clustered.png](http://www.progenetix.net/bcats/clustered.png)).



Michael Baudis  
BCATS Biocomputing at Stanford  
Stanford Nov 2001



## progenetix.net website from June 2003

**WAY BACK MACHINE - This is the Progenetix website version archived in 2003. Please go directly to [progenetix.org](http://progenetix.org) for some current data ...**

**[progenetix.net] molecular-cytogenetic data collection**

Please read the [license](#), especially if you are not from an academic institution.

**Collection of published cytogenetic abnormalities in human malignancies**  
For all cases registered in [progenetix], band specific chromosomal aberration data is available to be included in data mining projects. The complete dataset can be accessed for download (see [here](#) for information).

The **ISCN2matrix converter** allows the online conversion from an aberration list in ISCN format to a band specific aberration matrix, with optional generation of a graphical representation.

**Software source** for storage and visualization of CGH data

**7604 cases from 274 publications**  
Newest resolution: **863 bands**, matched to the "Golden Path" and **ENSEMBL CytoView**  
presented at **BCATS 2001** and **2002 (poster)** and the **ASH 2001** meeting

**Citation**

- Progenetix CGH online database. Baudis M. (2000-2003): [www.progenetix.net](http://www.progenetix.net)
- Progenetix.net: an online repository for molecular cytogenetic aberration data. Baudis M. and Cleary M. *Bioinformatics* 17 (12) 2001: 1228-1229.

**Submission**  
Casetables should be sent to [progenetix.net](mailto:progenetix.net).

**Server & Browser**  
The new version of the site is run on a commercial server, using RedHat Linux and Apache server software. It is optimized for newer generation browsers and is tested using Camino under OS X.

The page was generated at 5:7 (Pacific), 2003-6-17.



**Progenetix.net: an online repository for molecular cytogenetic aberration data**

Michael Baudis <sup>1,2,\*</sup> and Michael L. Cleary<sup>2</sup>

<sup>1</sup>Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and  
<sup>2</sup>Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001

### ABSTRACT

**Summary:** Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. [www.progenetix.net](http://www.progenetix.net) is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.

**Availability:** <http://www.progenetix.net>  
**Contact:** mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through <http://cgap.nci.nih.gov/Chromosomes/Mitelman>).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallion-

iemi et al., 1992; du Manoir et al., 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparison of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available<sup>1</sup>.

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

<sup>1</sup>Links to a number of online CGH resources with different scopes can be found at [www.progenetix.net](http://www.progenetix.net).

1228

© Oxford University Press 2001

Domain Name: PROGENETIX.NET  
Registry Domain ID: 45628826\_DOMAIN\_NET\_VRSN  
Registrar WHOIS Server: whois.enterprise.net  
Registrar URL: <http://www.epag.de>  
Updated Date: 2023-11-30T08:32:22Z  
Creation Date: 2000-11-29T18:17:38Z

## Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over **140'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services



### Cancer CNV Profiles

ICD-O Morphologies  
ICD-O Organ Sites  
Cancer Cell Lines  
Clinical Categories

### Search Samples

arrayMap  
TCGA Samples  
1000 Genomes  
Reference Samples  
DIPG Samples  
cBioPortal Studies  
Gao & Baudis, 2021

### Publication DB

Genome Profiling  
Progenetix Use

### Services

NCIt Mappings  
UBERON Mappings

### Upload & Plot

### Beacon<sup>+</sup>

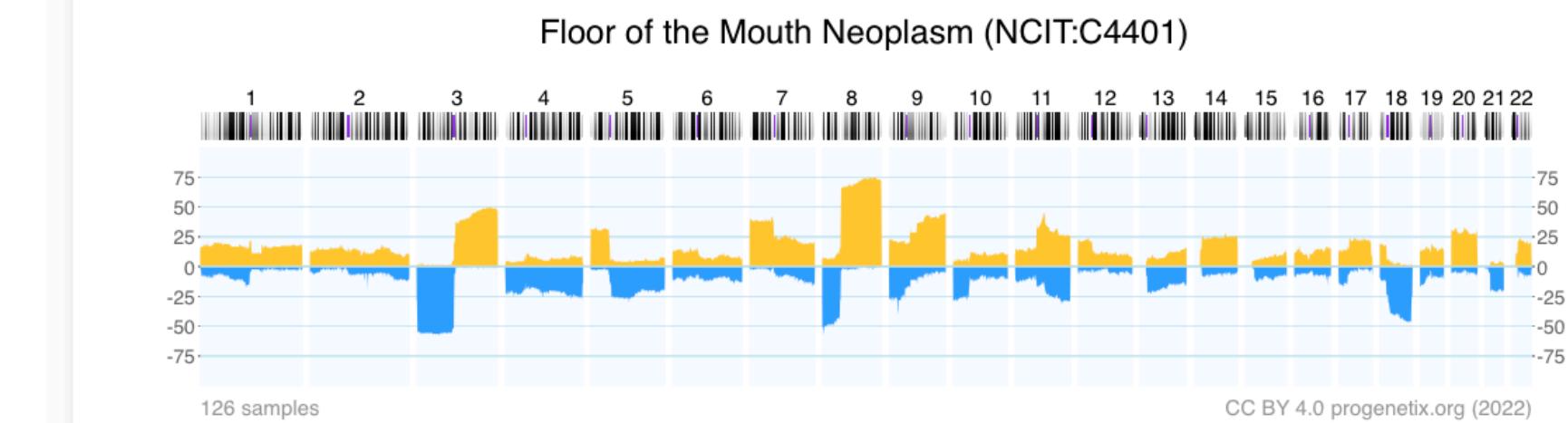
### Documentation

News  
Downloads & Use  
Cases  
Sevices & API

### Baudisgroup @ UZH

## Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.



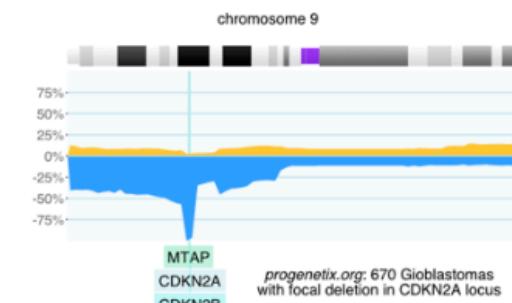
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.  
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

### Progenetix Use Cases

#### Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [\[ Search Page \]](#) provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.



progenetix.org: 670 Glioblastomas with local deletion in CDKN2A locus

#### Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [\[ Cancer Types \]](#) page with direct visualization and options for sample retrieval and plotting options.

#### Cancer Genomics Publications

Through the [\[ Publications \]](#) page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

# Standards Development & Implementation: CNV Terms

## in computational (file/schema) formats

- EFO:0030064
- EFO:0030067
  - | - EFO:0030068
  - \ - EFO:0020073
    - \ - EFO:0030069
- EFO:0030070
  - | - EFO:0030071
    - \ - EFO:0030072

<b>GA4GH VRS1.3+</b>	<b>Beacon v2</b>	<b>VCF v4.4</b>	<b>SO</b>
<b>EFO:0030070</b> gain	DUP or <b>EFO:0030070</b>	DUP SVCLAIM=D	SO:0001742 copy_number_gain
<b>EFO:0030071</b> low-level gain	DUP or <b>EFO:0030071</b>	DUP SVCLAIM=D	SO:0001742 copy_number_gain
<b>EFO:0030072</b> high-level gain	DUP or <b>EFO:0030072</b>	DUP SVCLAIM=D	SO:0001742 copy_number_gain
<b>EFO:0030072</b> high-level gain	DUP or EFO:0030073	DUP SVCLAIM=D	SO:0001742 copy_number_gain
<b>EFO:0030067</b> loss	DEL or <b>EFO:0030067</b>	DEL SVCLAIM=D	SO:0001743 copy_number_loss
<b>EFO:0030068</b> low-level loss	DEL or <b>EFO:0030068</b>	DEL SVCLAIM=D	SO:0001743 copy_number_loss
<b>EFO:0020073</b> high-level loss	DEL or <b>EFO:0020073</b>	DEL SVCLAIM=D	SO:0001743 copy_number_loss
<b>EFO:0030069</b> complete genomic loss	DEL or <b>EFO:0030069</b>	DEL SVCLAIM=D	SO:0001743 copy_number_loss

# Beacon v2 Filters

# **Example: Use of hierarchical classification systems (here NCI neoplasm core)**

- Beacon v2 relies heavily on "filters"
    - ontology term / CURIE
    - alphanumeric
    - custom
  - Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications
    - implicit *OR* with otherwise assumed *AND*
  - implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations



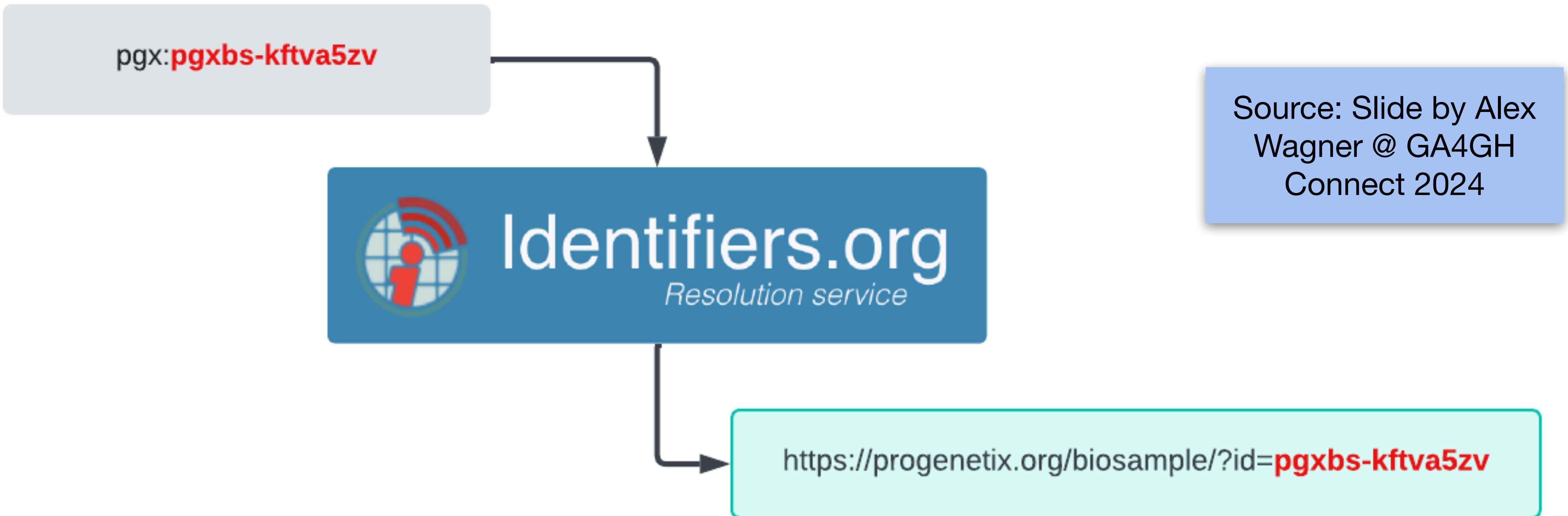
Beacon+ specific: Multiple term selection with OR logic

<input checked="" type="checkbox"/>	> <a href="#">NCIT:C4914: Skin Carcinoma</a>	213
<input type="checkbox"/>	> <a href="#">NCIT:C4475: Dermal Neoplasm</a>	109
<input checked="" type="checkbox"/>	> <a href="#">NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm</a>	310

**Filters:** NCIT:C4914, NCIT:C4819, NCIT:C9231, NCIT:C2921, NCIT:C45240, NCIT:C6858, NCIT:C3467, NCIT:C45340, NCIT:C7195, NCIT:C3246, NCIT:C7217

progenetix							
Variants: 0	$f_{alleles}$ : 0	<a href="#">Callsets</a>	<a href="#">Variants</a>	<a href="#">UCSC region</a>	<a href="#">Legacy Interface</a>	<a href="#"> Show JSON Response</a>	
Calls: 0							
Samples: 523							
Results	Biosamples						
Id	Description	Classifications		Identifiers	DEL	DUP	CNV
<a href="#">PGX_AM_BS_MCC01</a>	Merkel cell carcinoma	<a href="#">icdot-C44.9 Skin, NOS</a> <a href="#">icdom-82473 Merkel cell carcinoma</a> <a href="#">NCIT:C9231 Merkel Cell Carcinoma</a>		<a href="#">PMID:9537255</a>	0.116	0.104	0.22
<a href="#">PGX_AM_BS_MCC02</a>	Merkel cell carcinoma	<a href="#">icdot-C44.9 Skin, NOS</a> <a href="#">icdom-82473 Merkel cell carcinoma</a> <a href="#">NCIT:C9231 Merkel Cell Carcinoma</a>		<a href="#">PMID:9537255</a>	0.154	0.056	0.21
<a href="#">PGX_AM_BS_MCC03</a>	Merkel cell carcinoma	<a href="#">icdot-C44.9 Skin, NOS</a> <a href="#">icdom-82473 Merkel cell carcinoma</a> <a href="#">NCIT:C9231 Merkel Cell Carcinoma</a>		<a href="#">PMID:9537255</a>	0.137	0.21	0.347
<a href="#">PGX_AM_BS_MCC04</a>	Merkel cell carcinoma	<a href="#">icdot-C44.9 Skin, NOS</a> <a href="#">icdom-82473 Merkel cell carcinoma</a> <a href="#">NCIT:C9231 Merkel Cell Carcinoma</a>		<a href="#">PMID:9537255</a>	0.158	0.056	0.214
<a href="#">PGX_AM_BS_MCC05</a>	Merkel cell carcinoma	<a href="#">icdot-C44.9 Skin, NOS</a> <a href="#">icdom-82473 Merkel cell carcinoma</a> <a href="#">NCIT:C9231 Merkel Cell Carcinoma</a>		<a href="#">PMID:9537255</a>	0.107	0.327	0.434
<a href="#"></a>	<a href="#"></a>	<a href="#"></a>	<a href="#"></a>	Page 1 of 105			

## CURIE Resolution: Identifiers.org



# Progenetix as Genomics Resource

## Some trajectories ...

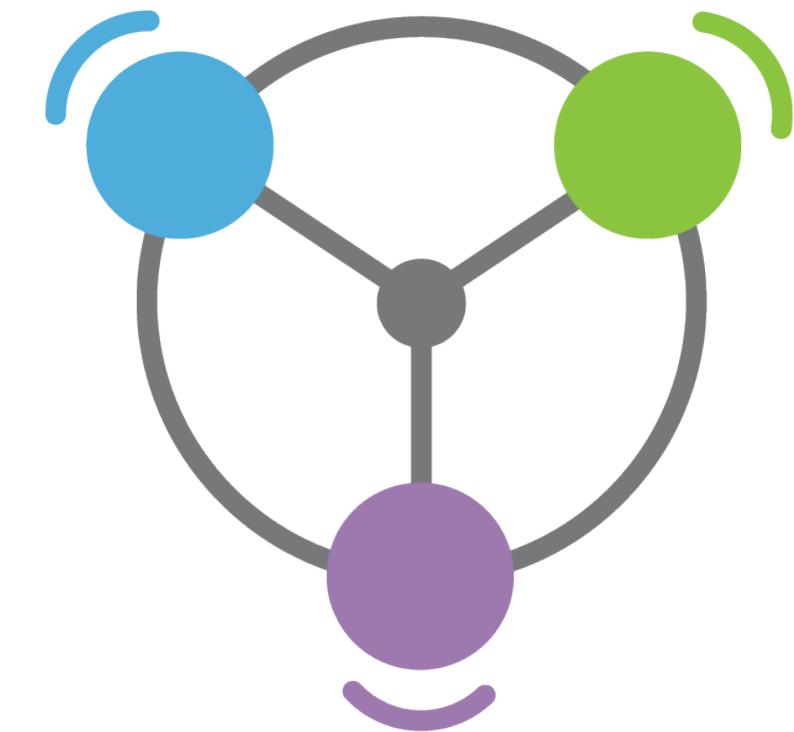
- from flat database to **hierarchical object storage**
- from dedicated database to mix of **open software tools**
- from static pages to **data driven website**
- from copy, paste, clean to **semi-automated download & process**
- from registering for raw data & commercial use to **CC BY 4.0 (CC0 for tools)**
- from local software development to **open code on Github**
- from standalone resource to federated data, **APIs** and services



# Different Approaches to Data Sharing



**Centralized Genomic Knowledge Bases**



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



**Linkage of distributed and disparate datasets**

# The EGA



Long term secure archive for human biomedical research sensitive data, with focus on reuse of the data for further research (or “*broad and responsible use of genomic data*”)



# The EGA

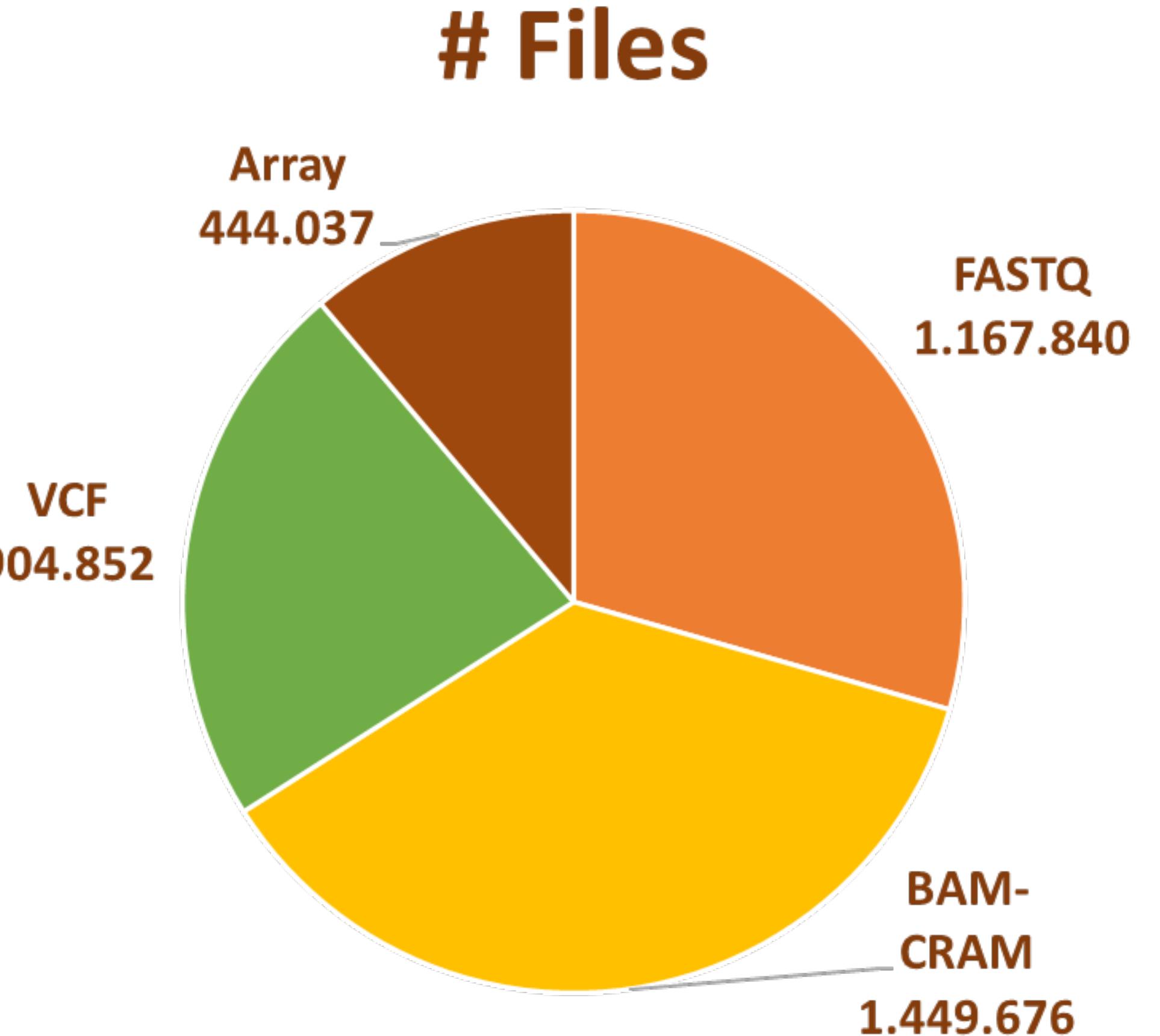


- EGA “owns” nothing; data controllers tell who is authorized to access ***their*** datasets
- EGA admins provide smooth “all or nothing” data sharing process

A screenshot of the EGA DAC interface. At the top left, it says "My DACs - EGAC5000000005 - Requests". Below that, it says "EuCanImage DAC" and "This is a DAC for EuCanImage data". On the right, there are "EDIT" and "HISTORY" buttons. The main area shows a "REQUESTS" section with a search bar and a table of recent requests. The table has columns for Date, Requester, Dataset, and DAC Admin/Member. The requests listed are:

Date	Requester	Dataset	DAC Admin/Member
18 August 2022	gemma.milla@crg.eu	EGAD5000000032	Dr Lauren A Fromont
17 August 2022	Dr Teresa Garcia Lezana	EGAD5000000033	Dr Teresa Garcia Lezana
16 August 2022	Dr Teresa Garcia Lezana	EGAD5000000032	Dr Lauren A Fromont

At the bottom right of the interface, there is an "APPLY" button.



4,328 Studies released  
10,470 Datasets  
2,309 Data Access Committees

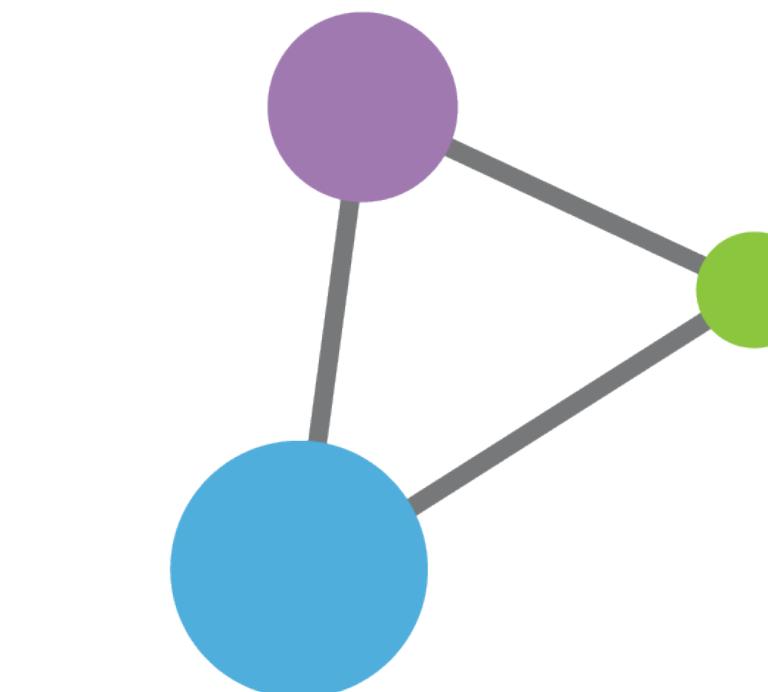
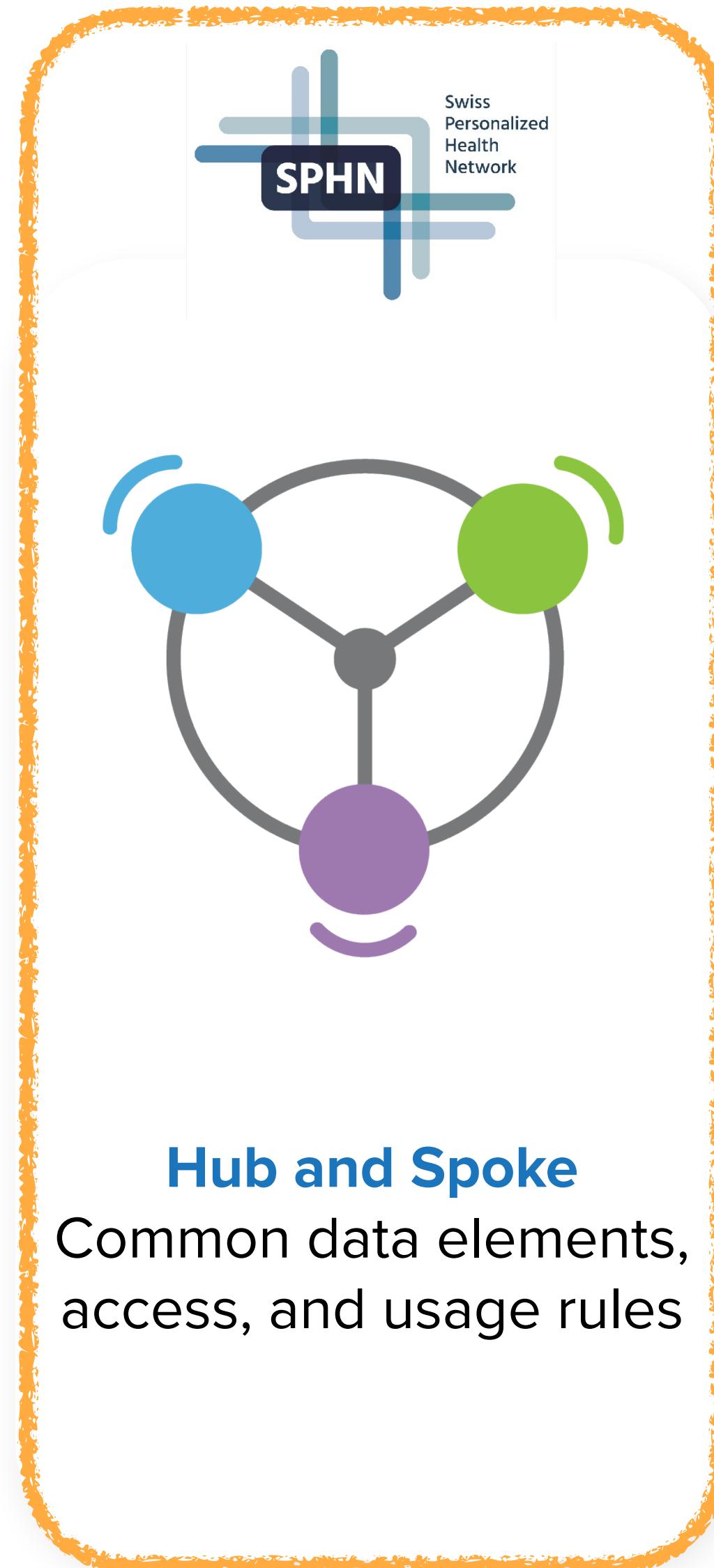
# Different Approaches to Data Sharing



**Centralized Genomic Knowledge Bases**

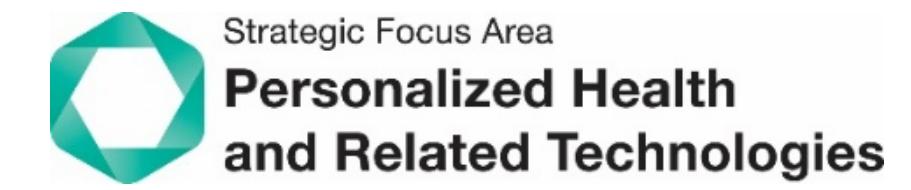
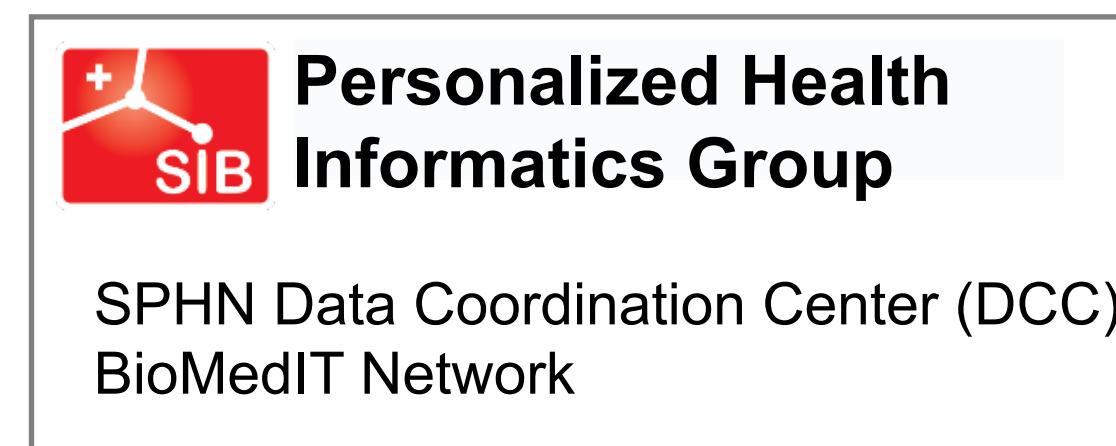
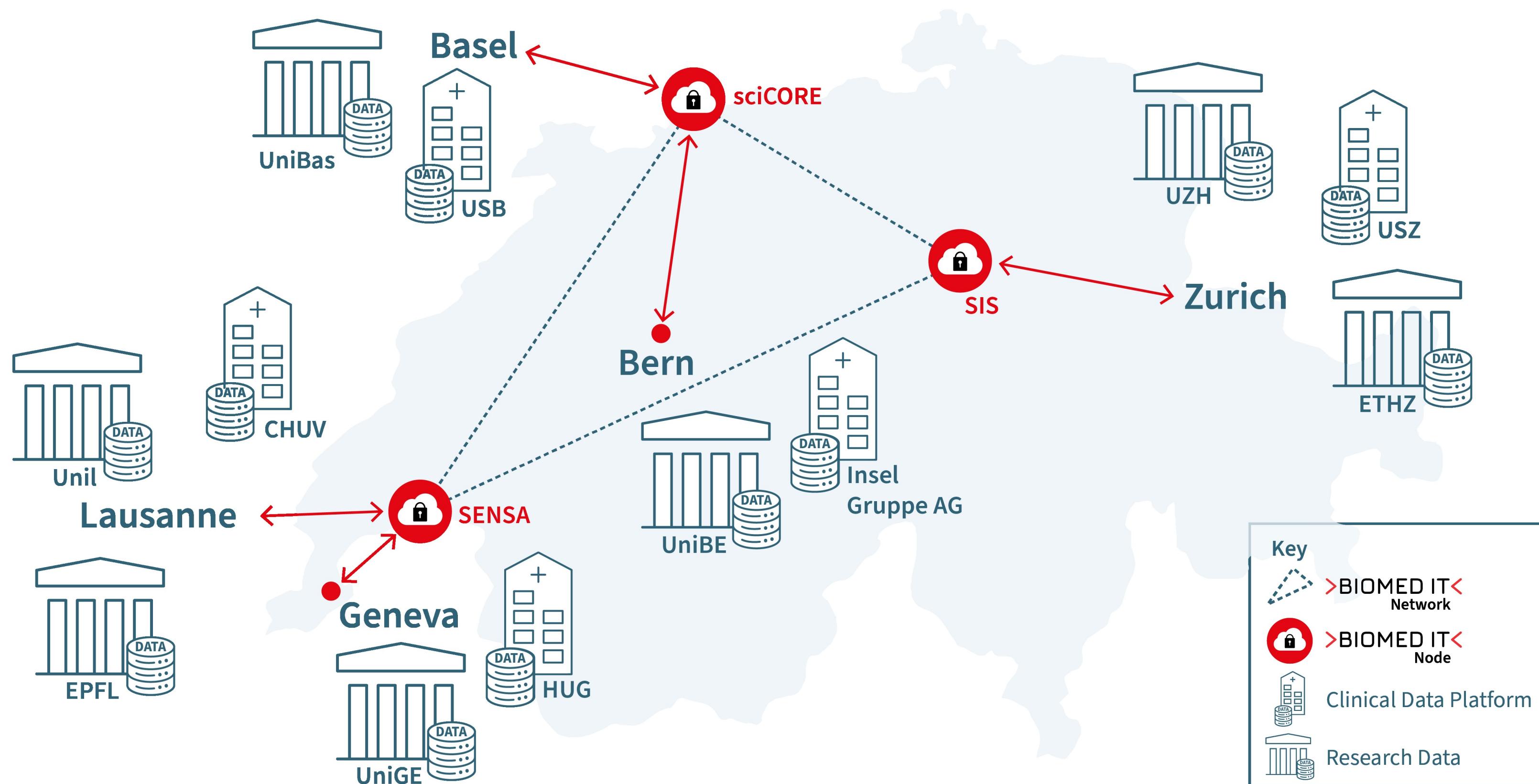


**Data Commons**  
Trusted, controlled repository of multiple datasets



**Linkage of distributed and disparate datasets**

# The Swiss Personalized Health Network



ehealthsuisse



Personalized Health Alliance  
Basel-Zurich



life sciences  
cluster basel



swissuniversities



# Different Approaches to Data Sharing



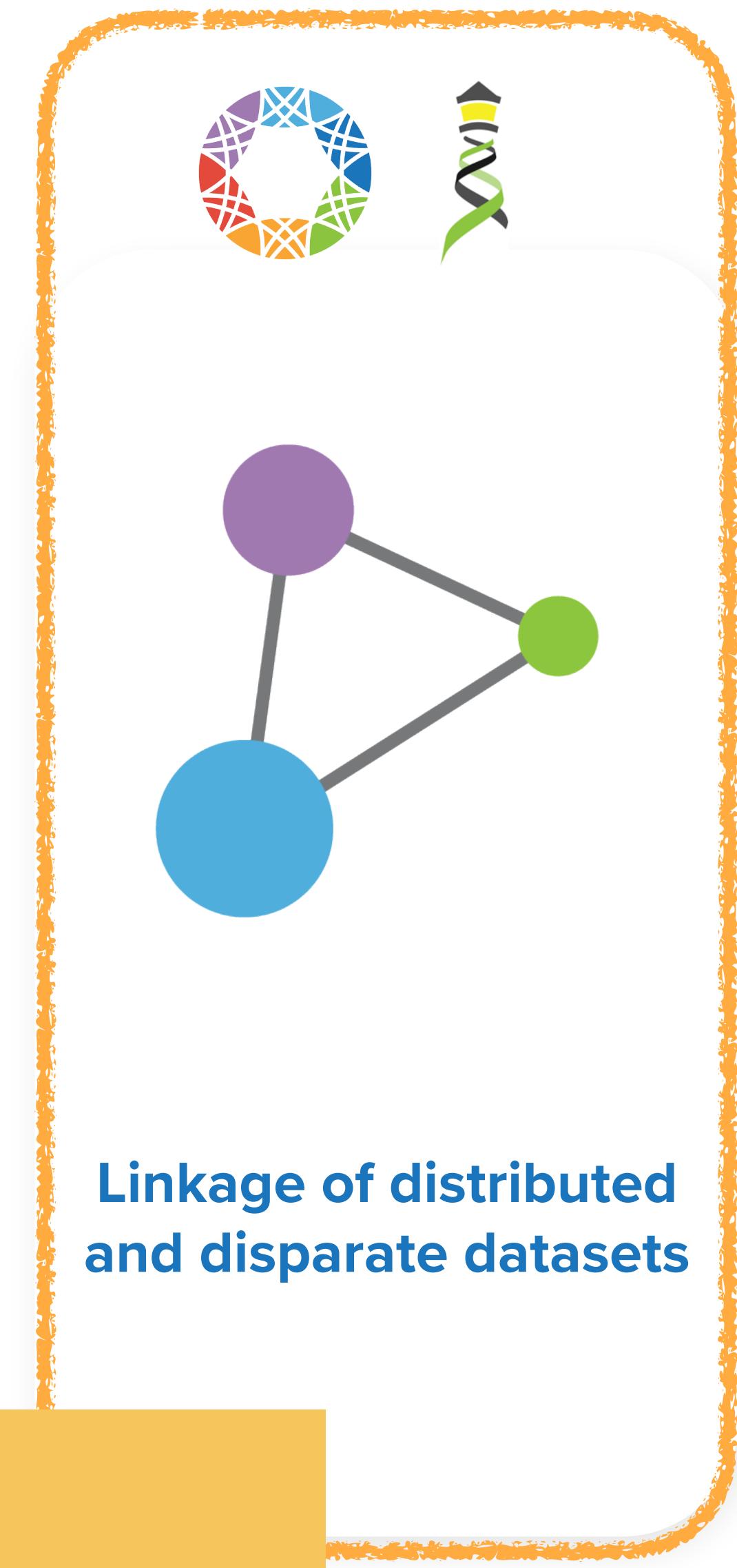
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

**Federation**

## INFORMATICS

### Beacon v2 and Beacon networks: federated data discovery in biome

#### Commentary

### International federation of genomic medicine databases using GA4GH standards

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

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

# Cell Genomics

## Technology

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

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

# Cell Genomics

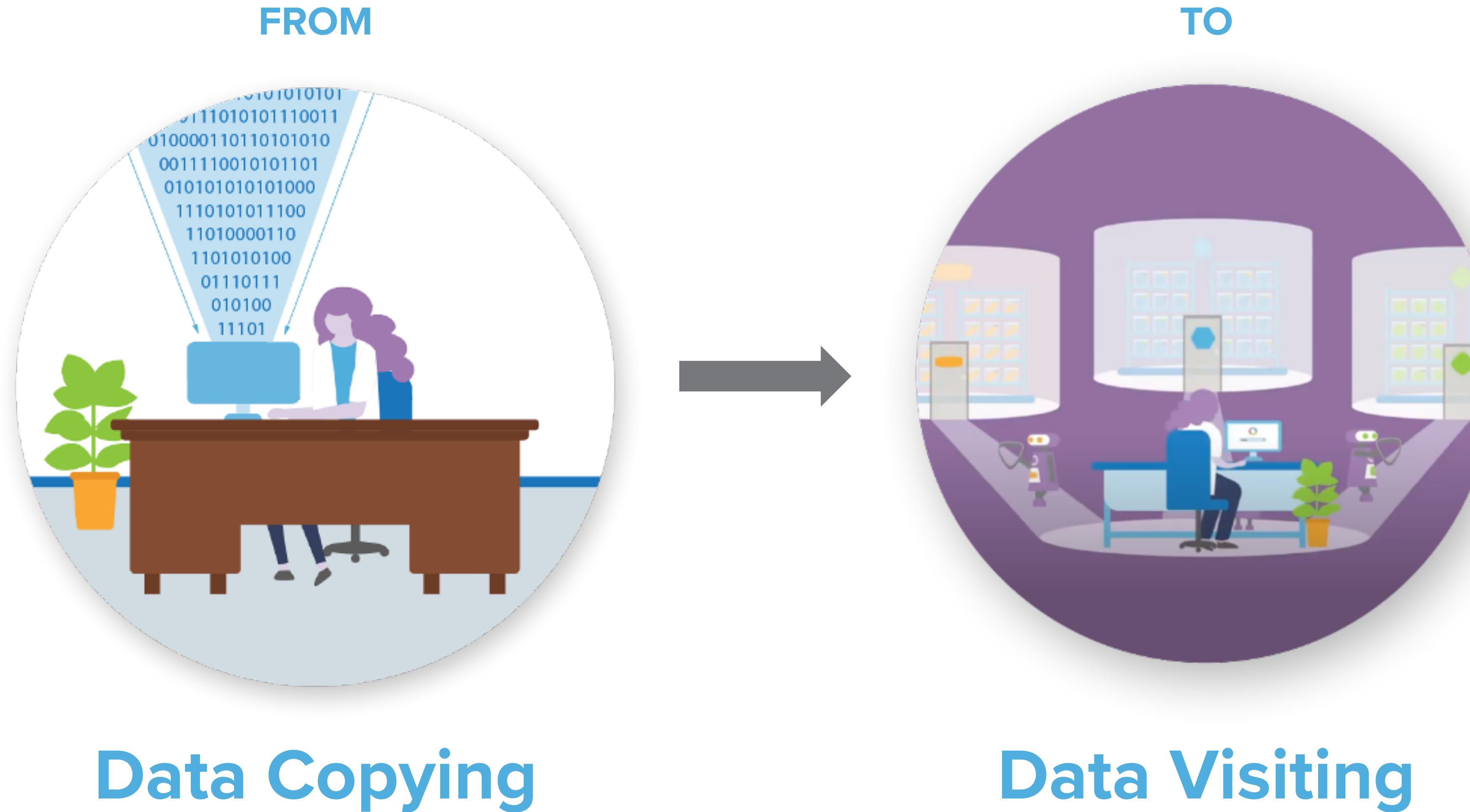
#### Perspective

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

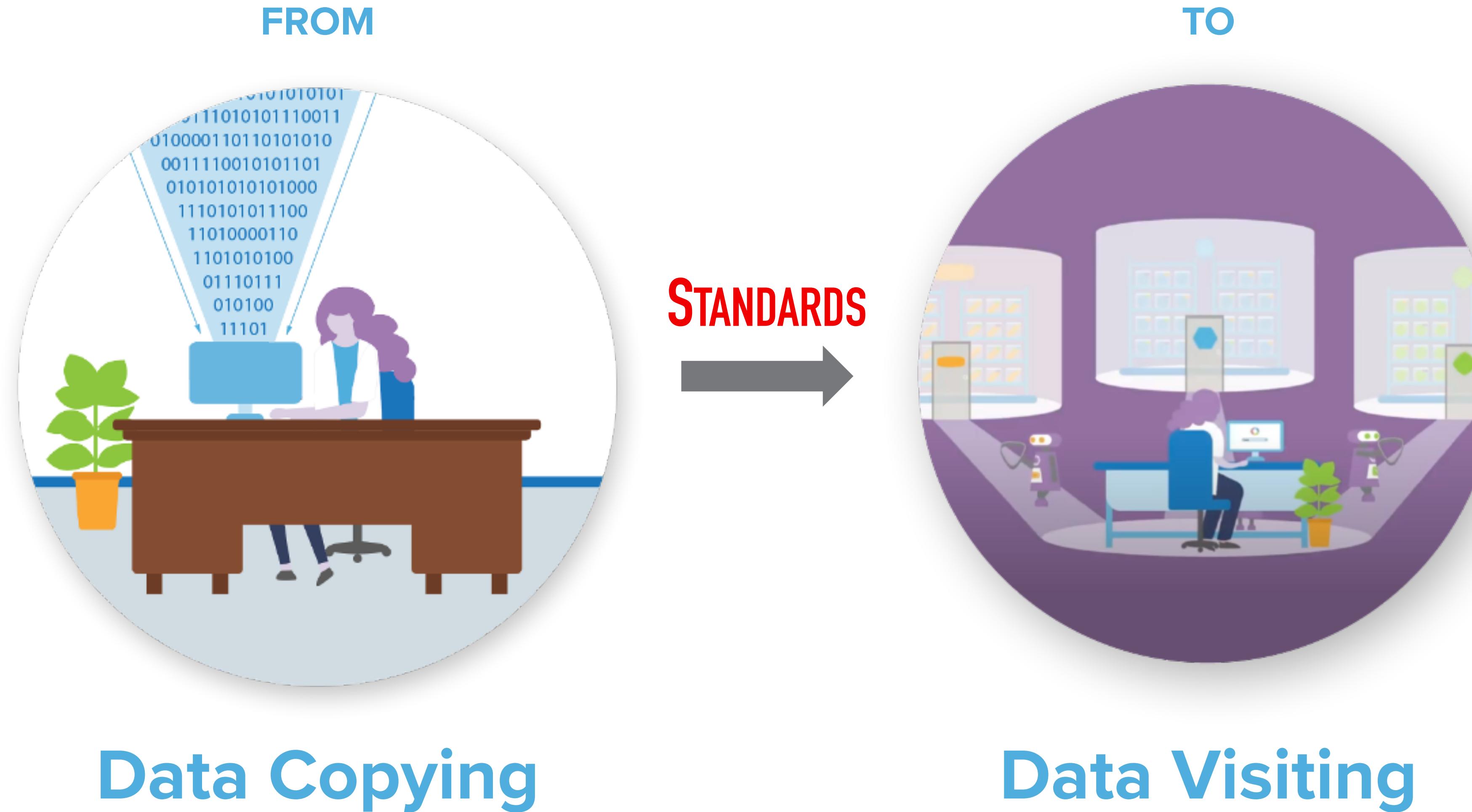
Heidi L. Rehm,<sup>1,2,47</sup> Angela J.H. Page,<sup>1,3,\*</sup> Lindsay Smith,<sup>3,4</sup> Jeremy B. Adams,<sup>3,4</sup> Gil Alterovitz,<sup>5,47</sup> Lawrence J. Babb,<sup>1</sup> Maxmillian P. Barkley,<sup>6</sup> Michael Baudis,<sup>7,8</sup> Michael J.S. Beauvais,<sup>3,9</sup> Tim Beck,<sup>10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

(Author list continued on next page)

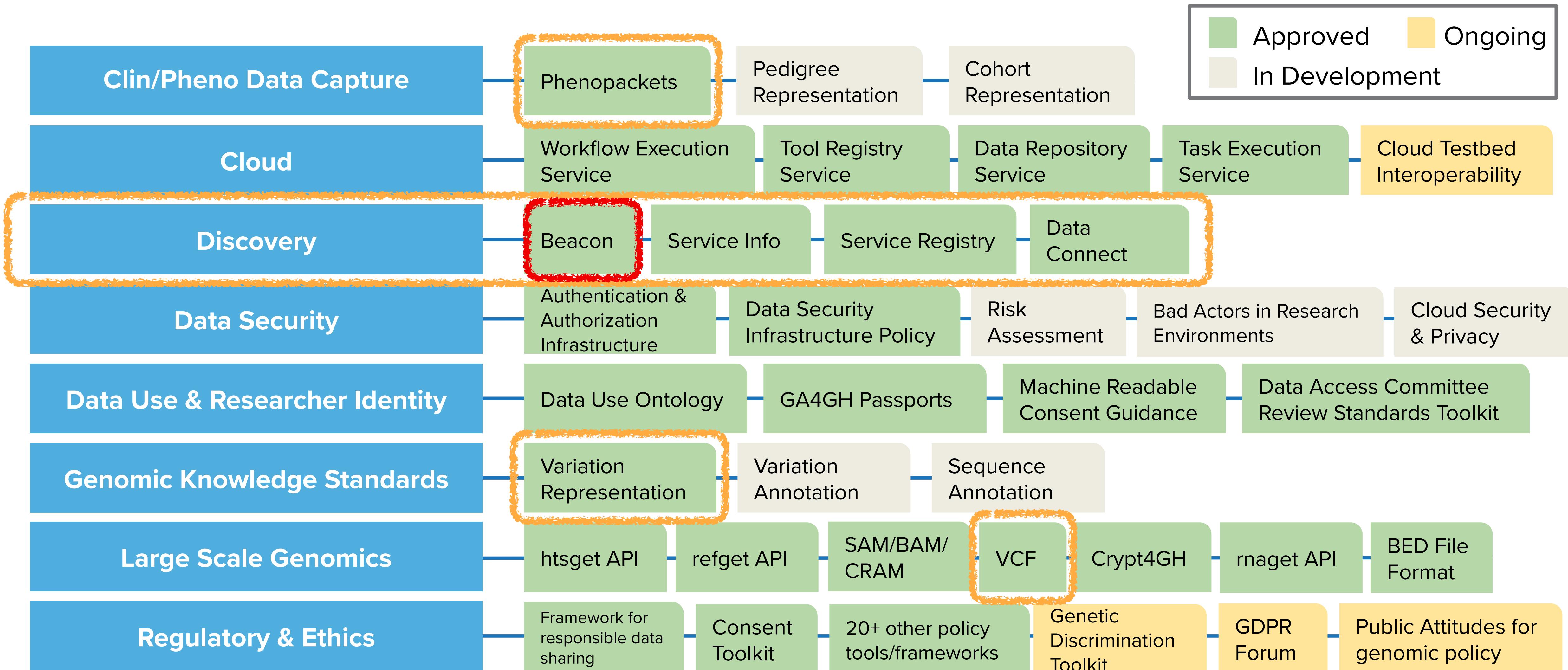
# A New Paradigm for Data Sharing



# A New Paradigm for Data Sharing



# Overview of GA4GH standards and frameworks



# Beacon v2 and Beyond

## The Standard for Data *Discovery* and Data *Sharing* in Biomedical Genomics



## Beacon v1 Development

2014

GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"

2015

- beacon-network.org aggregator created by DNAstack

2016

- Beacon v0.3 release
- work on queries for structural variants (brackets for fuzzy start and end parameters...)

2017

- OpenAPI implementation
- integrating **CNV parameters** (e.g. "startMin, statMax")
- Beacon v0.4 release in January; feature release for GA4GH approval process
- **GA4GH Beacon v1 approved** at Oct plenary

2019

- ELIXIR Beacon Network

2020



2021

2022

## Beacon v2 Development

- Beacon+ concept implemented @ [progenetix.org](#)
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")

- Beacon+ demos "handover" concept

- Beacon hackathon Stockholm; settling on **filters**
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept

- "**Scouts**" teams working on different aspects - filters, genomic variants, compliance ...

- **framework + models** concept implemented
- range and bracket queries, variant length
- starting of GA4GH review process

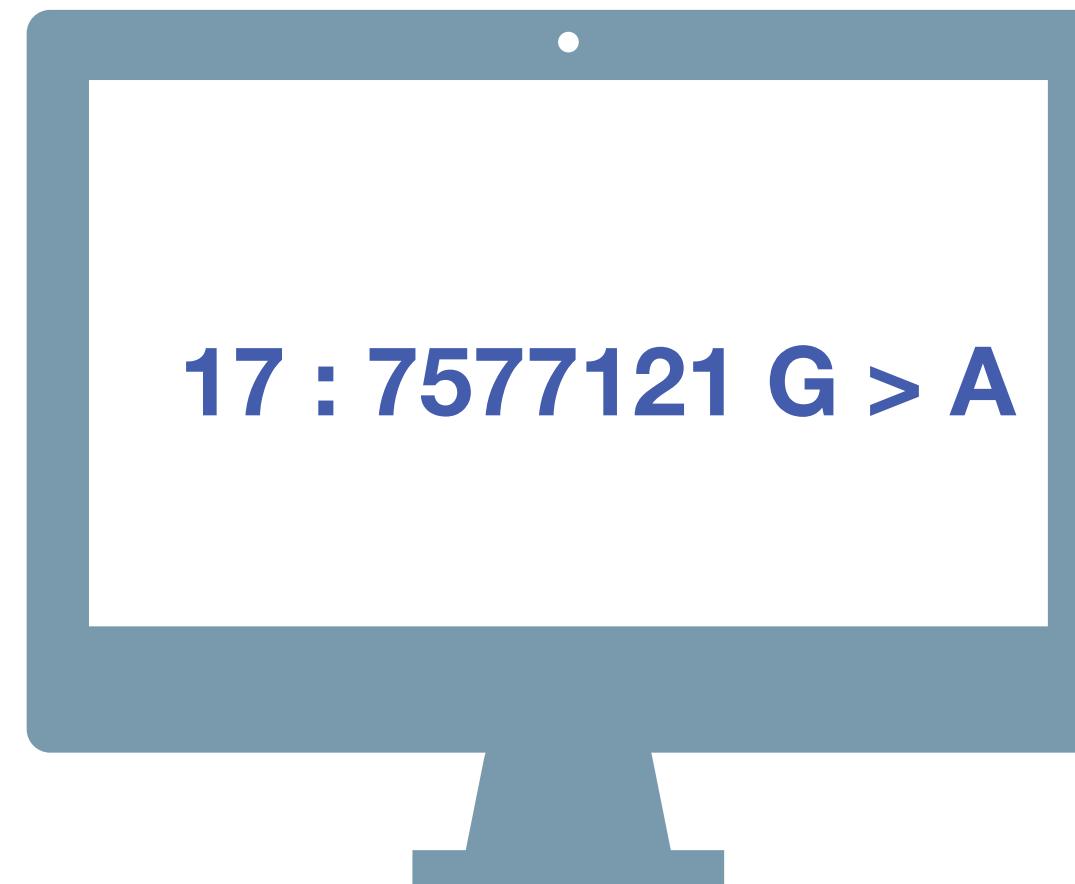
- changes in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- **Beacon v2 approved** at April GA4GH Connect

## Related ...

- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS
- new Beacon website (March)
- Beacon publication at Nature Biotechnology

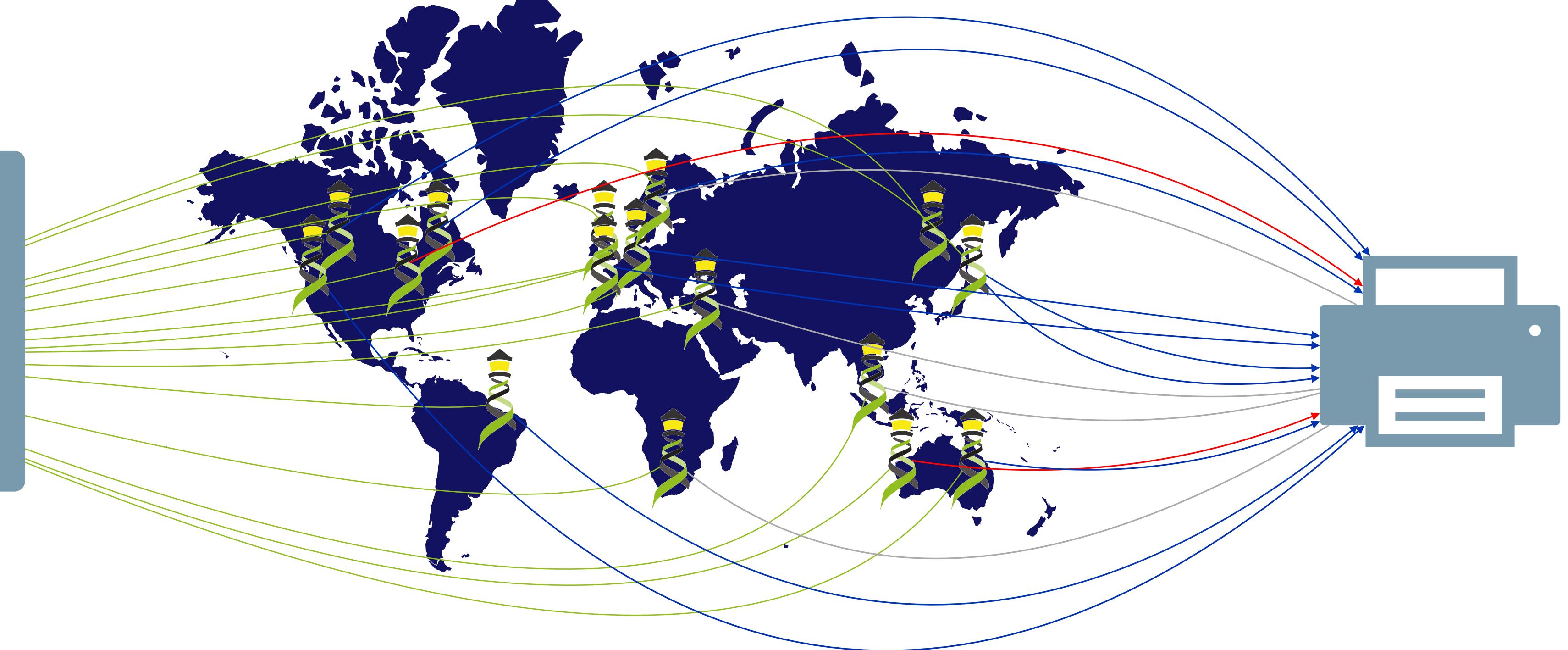
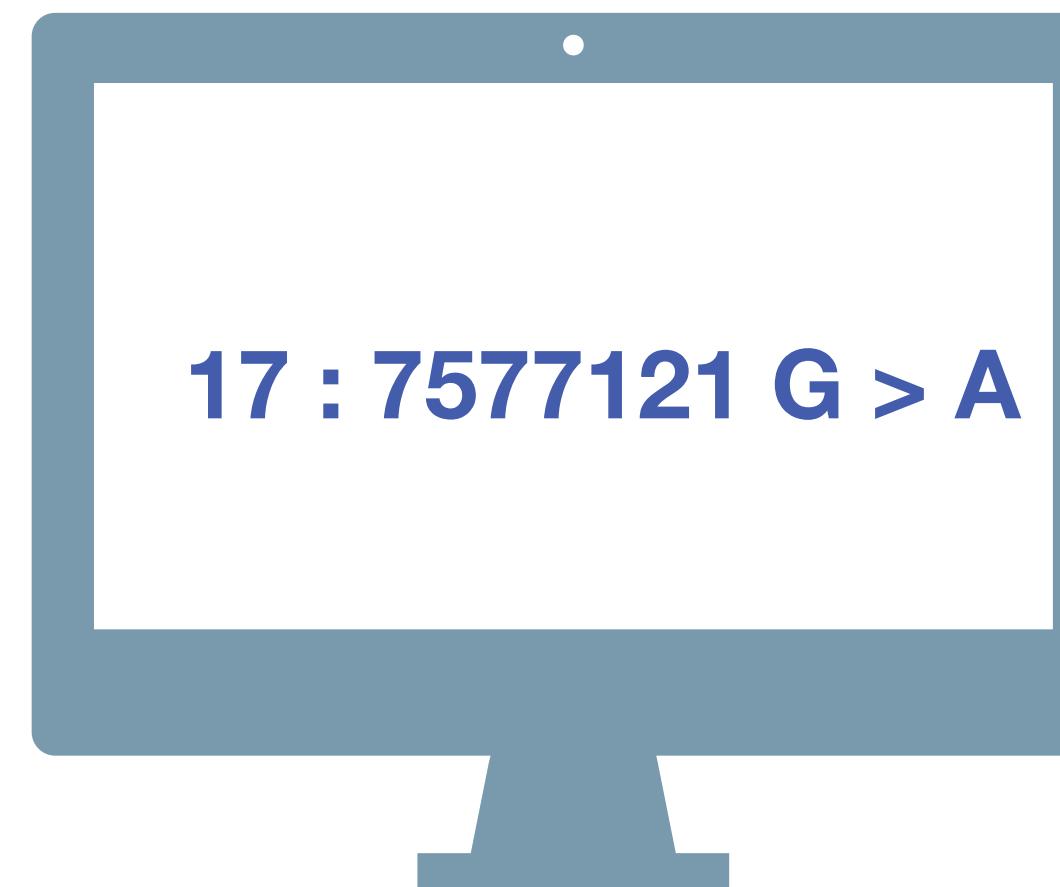
- Phenopackets v2 approved
- [docs.genomebeacons.org](#)



# Beacon

A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

**YES | NO | \0**



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**.

## Introduction

... I proposed a challenge application for all those wishing to seriously engage in *international* data sharing for human genomics. ...

1. Provide a public web service
2. Which accepts a query of the form “Do you have any genomes with an “A” at position 100,735 on chromosome 3?”
3. And responds with one of “Yes” or “No” ...

“Beacon” because ... people have been scanning the universe of human research for *signs of willing participants in far reaching data sharing*, but ... it has remained a dark and quiet place. The hope of this challenge is to 1) *trigger the issues* blocking groups ... in way that isn’t masked by the ... complexities of the science, fully functional interfaces, and real issues of privacy, and to 2) in *short order* ... see *real beacons of measurable signal* ... from *at least some sites* ... Once your “GABeacon” is shining, you can start to take the *next steps to add functionality* to it, and *finding the other groups* ... following their GABeacons.

## Utility

Some have argued that this simple example is not “useful” so nobody would build it. Of course it is not the first priority for this application to be scientifically useful. ...intended to provide a *low bar for the first step of real ... engagement*. ... there is some utility in ...locating a rare allele in your data, ... not zero.

A number of more useful first versions have been suggested.

1. Provide *frequencies of all alleles* at that point
2. Ask for all alleles seen in a gene *region* (and more elaborate versions of this)
3. Other more complicated queries

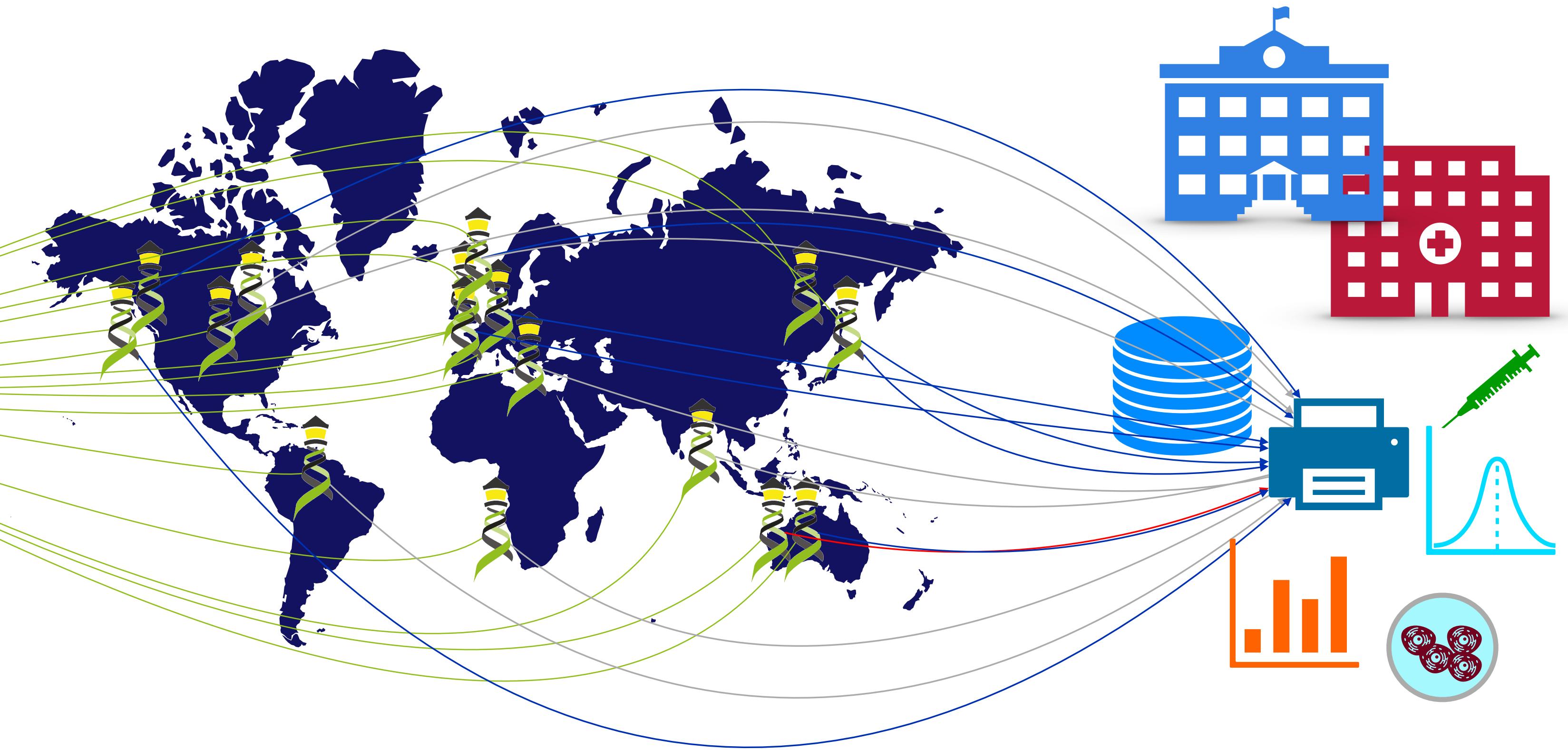
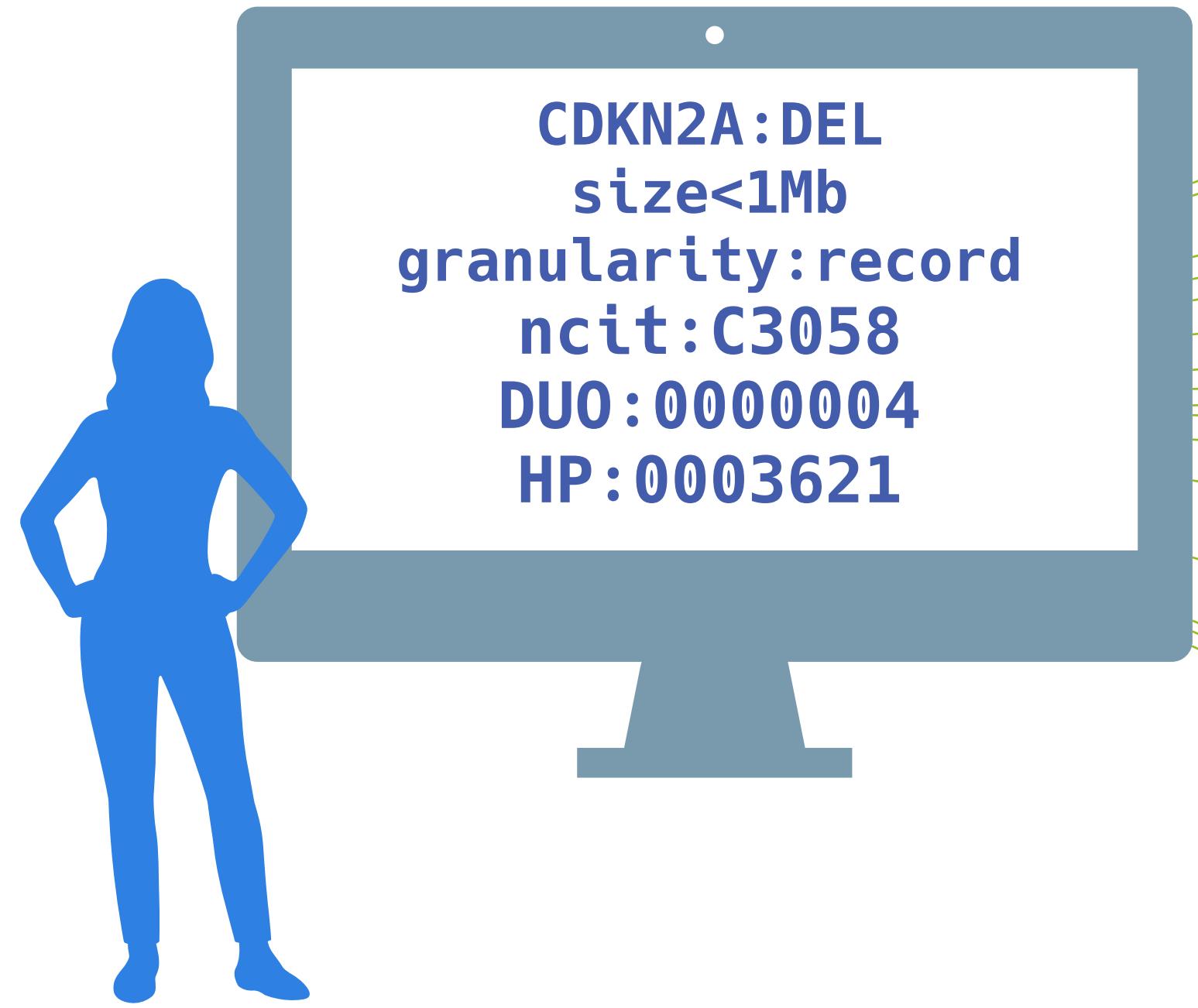


“I would personally recommend all those be held for  
**version 2**, when the beacon becomes a service.”

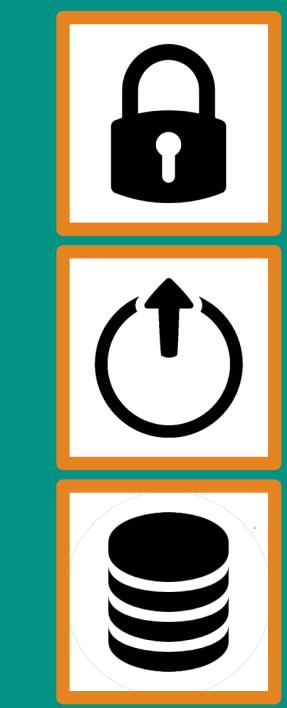
Jim Ostell, 2014

## Implementation

1. Specifying the chromosome ... The interface needs to specify the *accession.version* of a chromosome, or *build number*...
2. Return values ... right to *refuse* to answer without it being an error ... DOS *attack* ... or because ...especially *sensitive*...
3. Real time response ... Some sites suggest that it would be necessary to have a “*phone home*” *response* ...



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

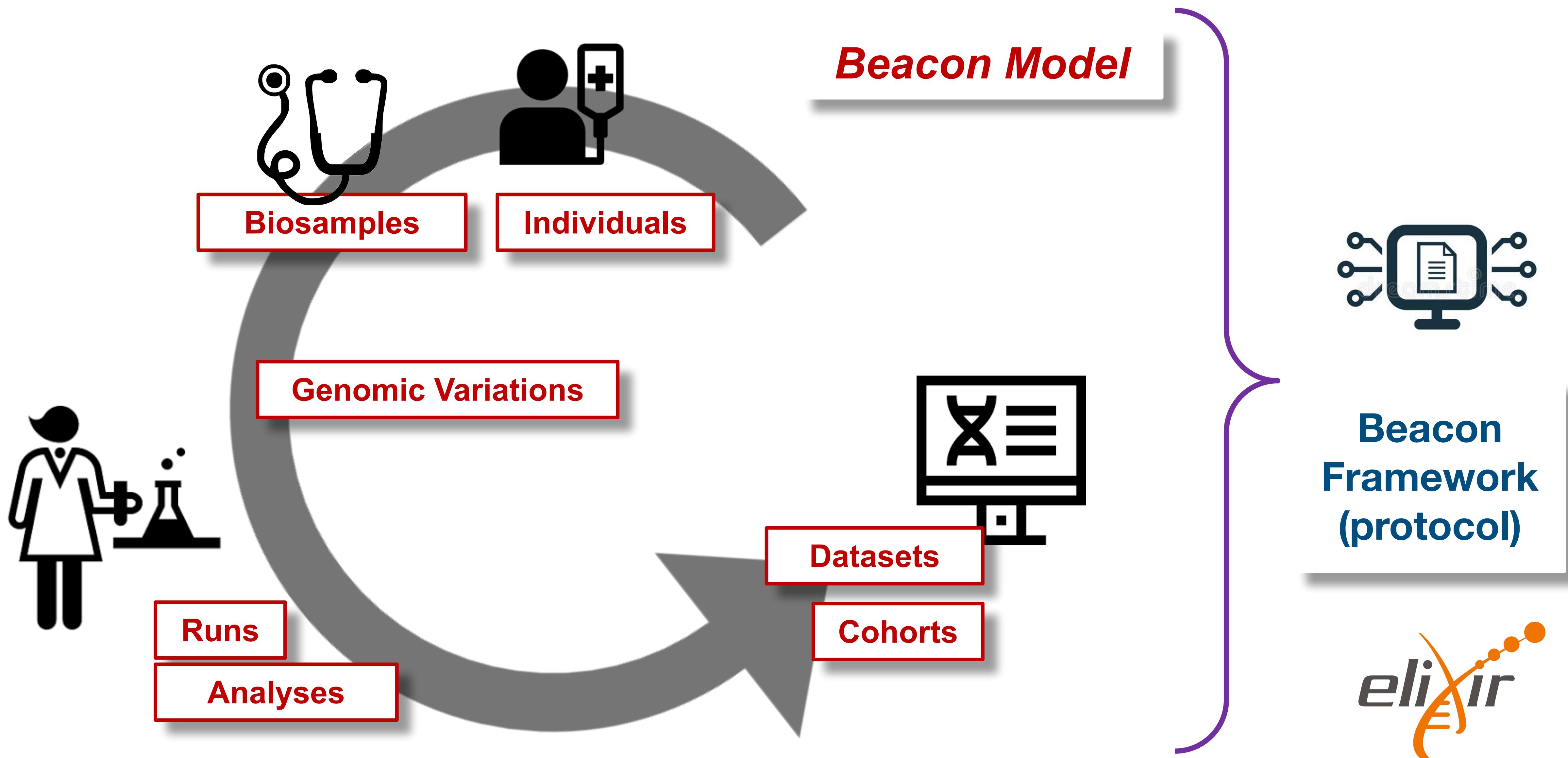


## Beacon **v2** API

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

# Beacon v2

docs.genomebeacons.org



**Beacon  
Framework  
(protocol)**



# Begriffsbestimmung

## The right expressions help to conceptualize...

- **Beacon:** The protocol/API, with framework and default model
- **beacon:** Implementation of Beacon
  - using the Beacon v2 framework & supporting at minimum boolean responses
  - suggested support of Beacon v2 default model but can choose other
- Beacon **Aggregator:** service distributes queries to beacons and aggregates responses into a single Beacon response
  - potential to liftover genomes, remap filtering terms, translate between protocol versions...
  - entry point to or potentially itself node in a ...
- Beacon **Network:** Set of beacons with shared entry point for distributed queries and aggregated response delivery
  - "true" beacon networks should have managed aspects - scope, term use...
  - networks may combine mixes of internal (protected, rich data, additional extensions...) and external interfaces

# bycon Beacon+

## Implementation driven standards development

- Progenetix' Beacon+ has served as implementation driver since 2016
- the *bycon* package is used to prototype advanced Beacon features such as
  - structural variant queries
  - data handovers
  - Phenopackets integration
  - variant co-occurrences
  - ...

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon v2 GA4GH Approval Registry

Beacons: European Genome-Phenome Archive | progenetix | cnag | University of Leicester

**European Genome-Phenome Archive (EGA)**

GA4GH Approval Beacon Test

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	Green
Bioinformatics analysis	Green
Biological Sample	Green
Cohort	Green
Configuration	Green
Dataset	Green
EntryTypes	Green
Genomic Variants	Green
Individual	Green
Info	Green
Sequencing run	Green

**progenetix**

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Progenetix Cancer Genomics Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

Visit us	Green
Beacon UI	Green
Beacon API	Green
Contact us	Green
BeaconMap	Green
Bioinformatics analysis	Green
Biological Sample	Green
Cohort	Green
Configuration	Green
Dataset	Green
EntryTypes	Green
Genomic Variants	Green
Individual	Green
Info	Green
Sequencing run	Green

**cnag**

Centre Nacional Analisis Genomica (CNAG-CRG)

Beacon @ RD-Connect

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	Green
Bioinformatics analysis	White
Biological Sample	Red
Cohort	Green
Configuration	Green
Dataset	Red
EntryTypes	Green
Genomic Variants	White
Individual	Red
Info	Red
Sequencing run	White

**University of Leicester**

Cafe Variome Beacon v2

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	Green
Bioinformatics analysis	White
Biological Sample	White
Cohort	White
Configuration	Green
Dataset	Green
EntryTypes	Green
Genomic Variants	Green
Individual	White
Info	Green
Sequencing run	White

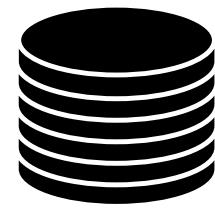
✓ Matches the Spec   ✗ Not Match the Spec   ● Not Implemented



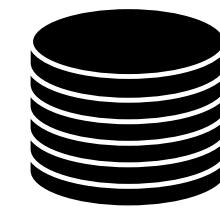
# *bycon* based Beacon+ Stack

progenetix

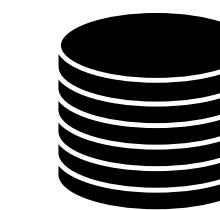
- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - ▶ [pubmed:10027410](#), [NCIT:C3222](#), [pgx:cohort-TCGA](#), [pgx:icdom-94703...](#)
  - ▶ precomputed frequencies per collection informative e.g. in form autfills
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding **accessid** for **handover** generation
- complete query aggregation; i.e. individual queries are run against the corresponding entities and ids are intersected
  - retrieval of any entity, e.g. all individuals which have queried variants analyzed on a given platform
  - allows multi-variant queries, i.e. all bio samples or individuals which had matches of all of the individual variant queries



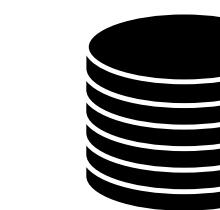
variants



analyses



biosamples



individuals



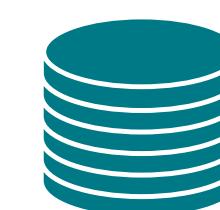
collations



geolocs



genespans

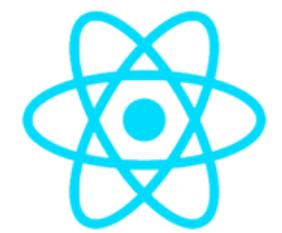


qBuffer

Entity collections

Utility collections

[github.com/progenetix/bycon](https://github.com/progenetix/bycon)



React



APACHE  
HTTP SERVER PROJECT



python™



mongoDB



# pgxRpi: an R/Bioconductor package

## Client for Accessing Beaconized Data

- **Query and export variants**

[https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g\\_variants](https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants)

```
> variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
```

- **Query metadata of biosamples and individuals by filters (e.g. NCIt, PMID)**

<http://progenetix.org/services/sampletable/?filters=NCIT:C3697>

```
> biosamples <- pgxLoader(type="biosample",filters="NCIT:C3697")
```

- **Query and visualize CNV frequency by filters**

<http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C3512>

```
> freq <- pgxLoader(type="frequency",output="pgxfreq",filter  
> pgxFreqplot(freq)
```

- **Process local .pgxseg files**

```
> info <- pgxSegprocess(file=file, show_KM_plot = T,  
return_seg = T, return_metadata = T, return_frequency = T)
```

## pgxRpi

This is the **development** version of pgxRpi; for the stable release version, see [pgxRpi](#).

### R wrapper for Progenetix

platforms all rank 2178 / 2266 support 0 / 0 in Bioc < 6 months build unknown updated < 1 month dependencies 137

DOI: [10.18129/B9.bioc.pgxRpi](https://doi.org/10.18129/B9.bioc.pgxRpi)

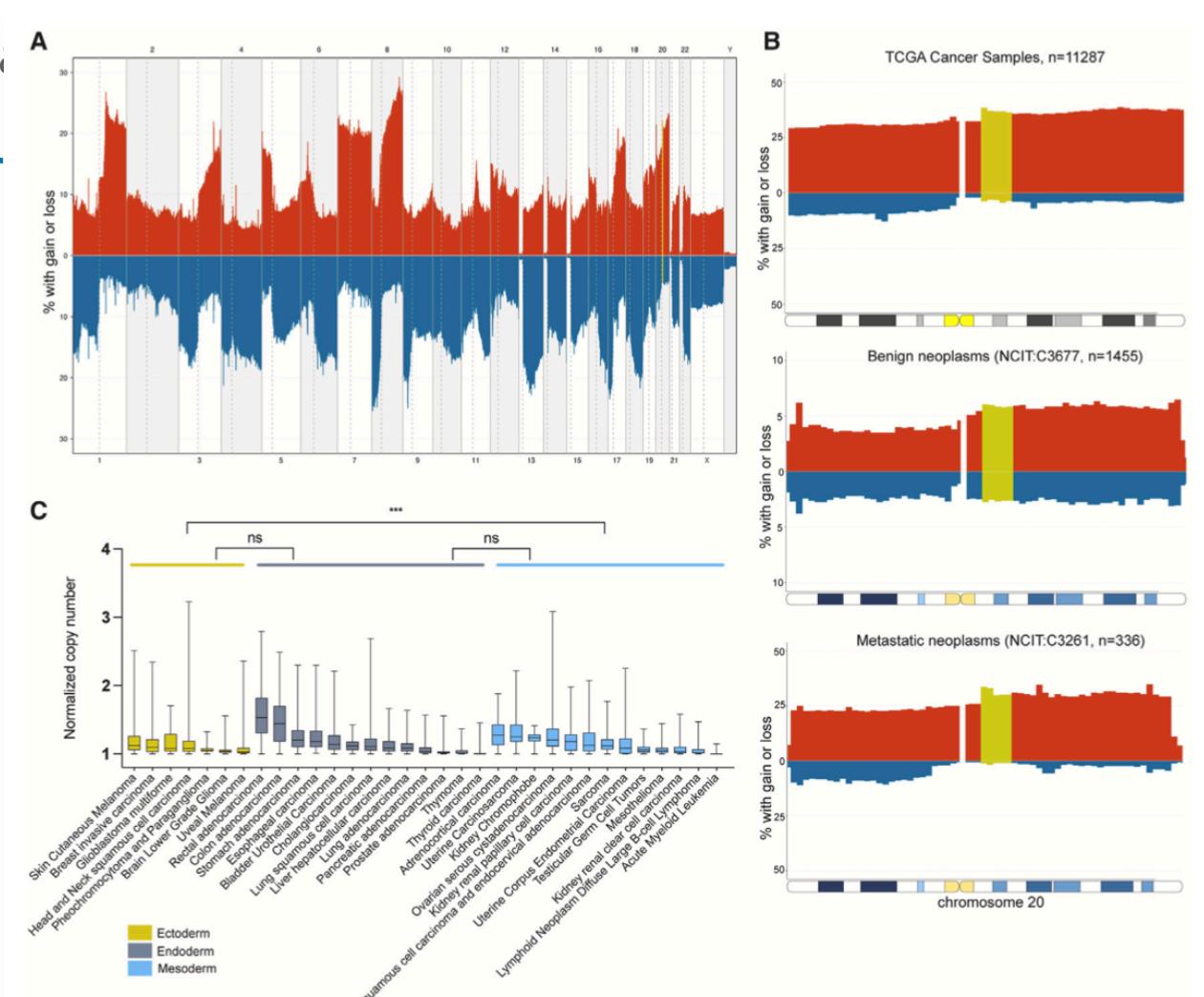
**Bioconductor version:** Development (3.20)

The package is an R wrapper for Progenetix REST API built upon the Beacon v2 protocol. Its purpose is to provide a seamless way for retrieving genomic data from Progenetix database—an open resource dedicated to curated oncogenomic profiles. Empowered by this package, users can effortlessly access and visualize data from Progenetix.

**Author:** Hangjia Zhao [aut, cre]  Michael Baudis [aut] 

**Maintainer:** Hangjia Zhao <hangjia.zhao at uzh.ch>

Use case: 2024 article using Progenetix' *pgxRpi* to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics



## Stem Cell Reports Review



OPEN ACCESS

### Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,<sup>1,2</sup> Manjusha S. Ghosh,<sup>1,2</sup> and Claudia Spits<sup>1,2,\*</sup>

<sup>1</sup>Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

<sup>2</sup>These authors contributed equally

\*Correspondence: [claudia.spits@vub.be](mailto:claudia.spits@vub.be)

<https://doi.org/10.1016/j.stemcr.2023.11.013>

#### Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers

(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.

(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

# **Beacon Security**



# Making Beacons Biomedical - Beacon v2

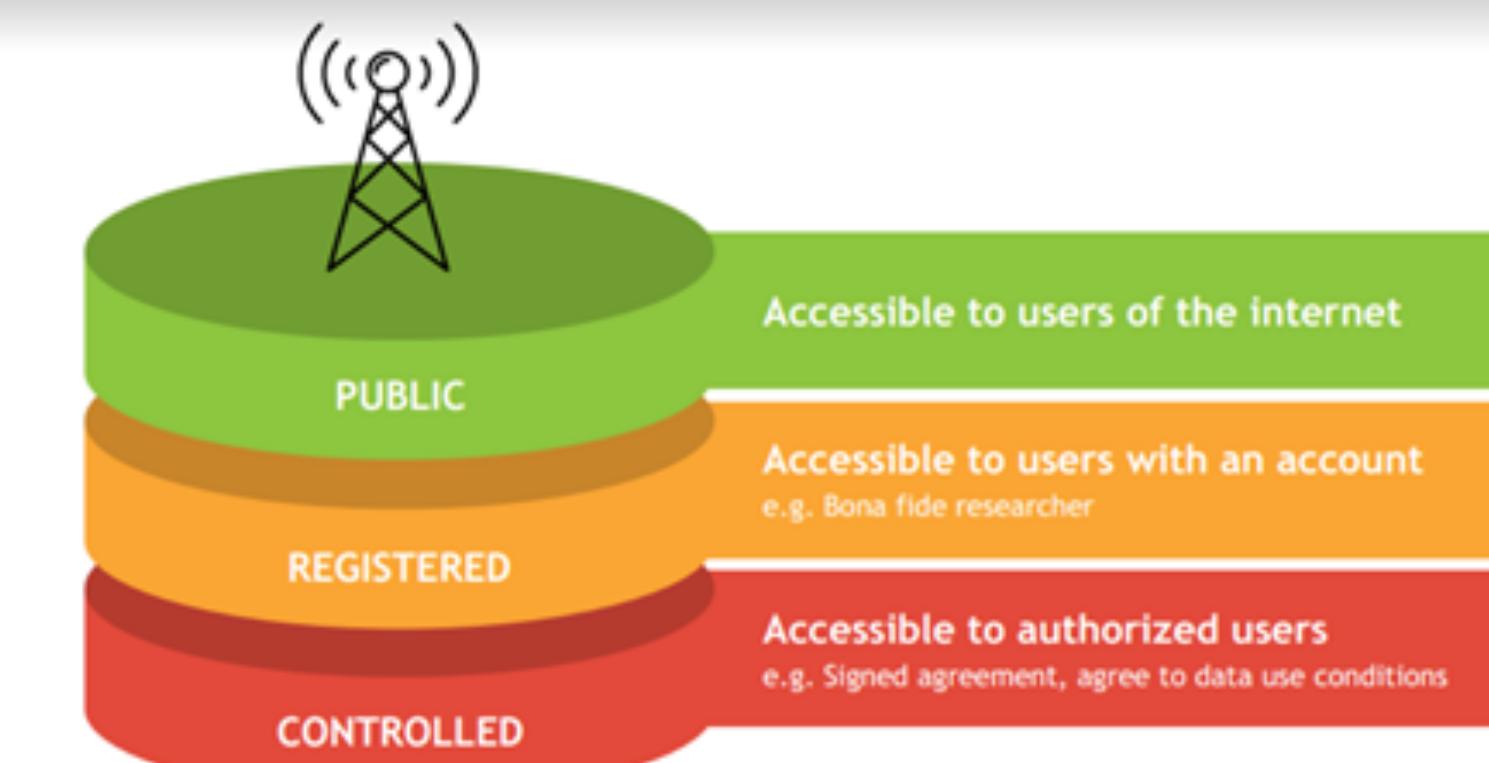
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant effects
- Beacon queries as entry for **data delivery**
  - Beacon v2 permissive to respond with variety of data types
    - Phenopackets, biosample data, cohort information ...
  - handover to stream and download using htsget, VCF, EHRs
- Interacting with EHR standards
  - FHIR translations for queries and handover ...
- Beacons as part of local, secure environments
- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)

Definitely breaks the  
"Relative Security  
by Design"  
Concept!

# Beacon Security

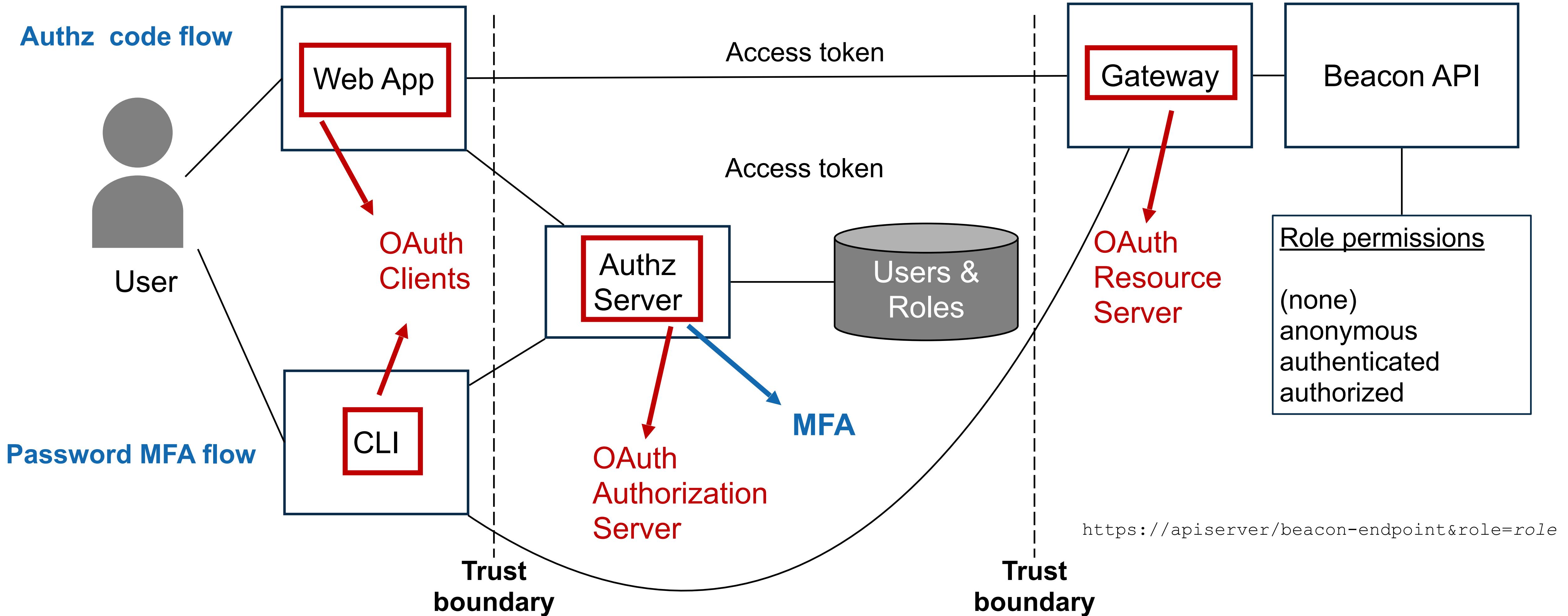
## Security by Design ... if Implemented in the Environment

- the beacon API specification does not implement explicit security (e.g. checking user authentication and authorization)
- the framework implements different levels of response granularity which can be mapped to authorization levels (**boolean** / **count** / **record** level responses)
- implementations can have beacons running in secure environments with a **gatekeeper** service managing authentication and authorization levels, and potentially can filter responses for escalated levels
- the backend can implement additional access reduction, on a user <-> dataset level if needed



# Architecture

## Running the *bycon* stack in a secure environment



# Architecture

## Running the *bycon* stack in a secure environment

- The **Beacon API** implementation stack (e.g. bycon) is authentication procedure agnostic; i.e. it just accepts that a user has been authenticated and passed the general authorization gatekeeping
- The **Beacon API** server and the **Gateway** reside in a single VM, with only the **Gateway**'s port exposed (with TLS). Beacon's port is not exposed by the VM and can only be reached through the **Gateway**
- The **Authentication Server** can run on the same or separate VM; needs a database with user accounts.
- The **Web Client** can be in the same VM or a separate one.
- Separate **Gateways** (e.g. university firewall vs. public) can be configured to modify different roles, e.g. the public gateway may turn registered roles into anonymous, regardless of whether the user has registered status
- Users can write their own clients (web / command line) which are registered with the **Authorization Server** and are issued with a Client ID and Client Secret to use against the **Authorization Server**.

**... to be continued**