Thursday 2025-04-17	Tuesday 2025-04-29	Wednesday 2025-04-30	Friday 2025-05-02	Tuesday 2025-05-06	Wednesday 2025-05-07
Michael: Survival Intro		Feifei & Ziying: survival	Feifei: population structure		Exam
Michael: Cancer Classifications		Feifei & Ziying: survival	Feifei: population structure		
Michael: Genomic Data Exchange & GA4GH		Feifei: survival	Feifei: population structure		
	Feifei:: analysis & interpretation. Parsing VCF (cyvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Feifei: population structure	Feifei & others: Presentation & Discussion	Michael: Genomic Data & Privacy	
	Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	Michael: Genomic Data & Privacy	
	Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	Discussion	

University of Zurich[UZH]

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

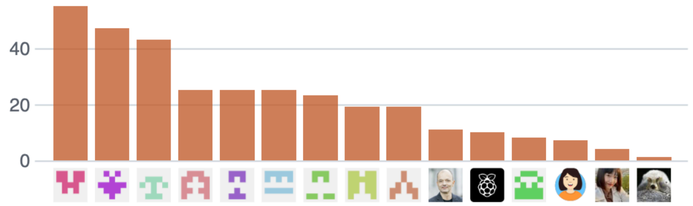
Overview

0 Active pull requests

0 Active issues

0 Merged pull requests	**0** Open pull requests	**0** Closed issues	**0** New issues

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-BIO392
➜  UZH-BIO392 git:(master)  mkdocs serve
```

Survival

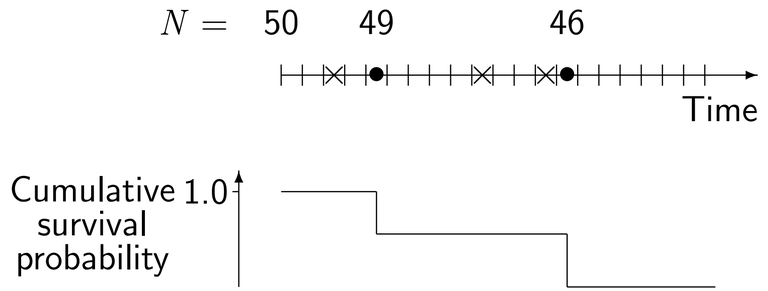
Kaplan-Meier Analysis of Survival Based on Conditional Probabilities

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

($\bullet$ = failure and $\times$ = censored):



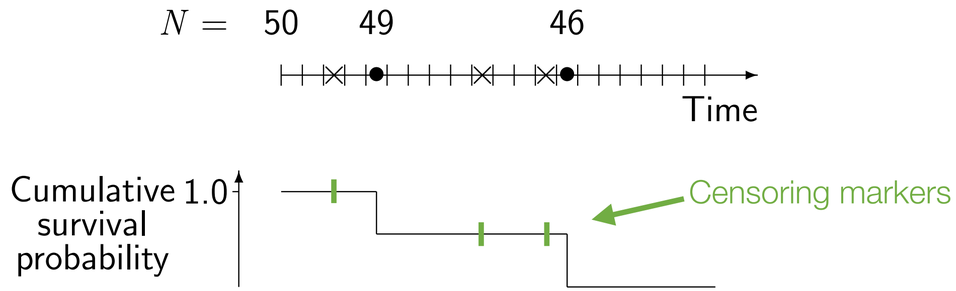
- Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively
- Late entry can also be dealt with

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

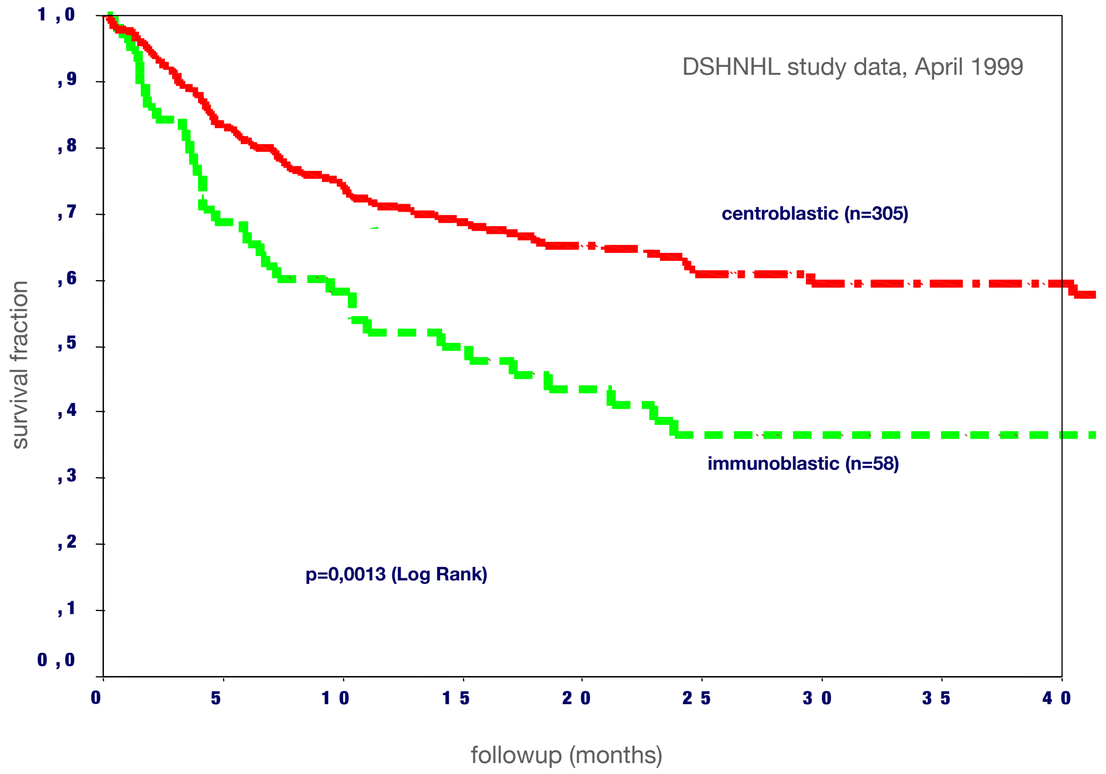
($\bullet$ = failure and $\times$ = censored):

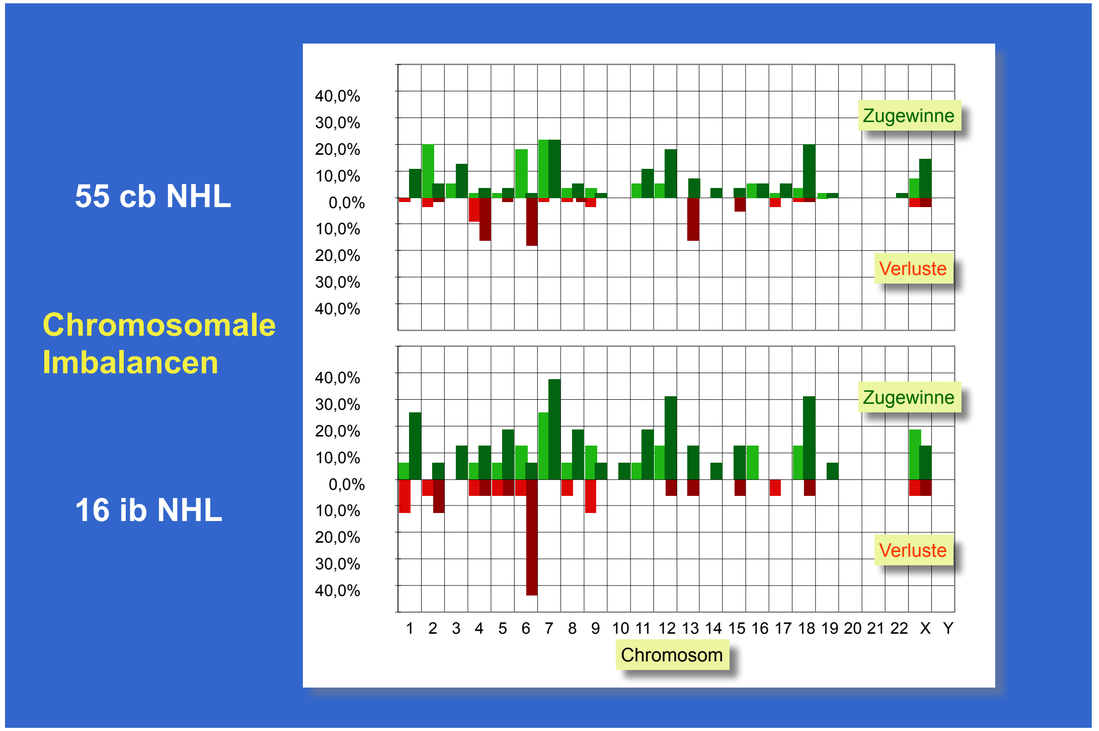


- Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively
- Late entry can also be dealt with

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. Hoechstetter[1] · Raymonde Busch[2] · Barbara Eichhorst[3] · Andreas Bühler[4] · Dirk Winkler[4] · Jasmin Bahlo[3] · Sandra Robrecht[3] · Michael J. Eckart[5] · Ursula Vehling-Kaiser[6] · Georg Jacobs[7] · Ulrich Jäger[8] · Hans Jürgen Hurtz[9] · Georg Hopfinger[10] · Frank Hartmann[11] · Harald Fuss[12] · Wolfgang Abenhardt[13] · Ilona Blau[14] · Werner Freier[15] · Lothar Müller[16] · Maria Goebeler[17] · Clemens Wendtner[1,3] · Kirsten Fischer[3] · Carmen D. Herling[3] · Michael Starck[1] · Martin Bentz[18] · Bertold Emmerich[19] · Hartmut Döhner[20] · Stephan Stilgenbauer[20] · Michael Hallek[3]

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		*P* value
			Lower	Upper	
COX regression OS					
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
COX regression TTFT					
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.**	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.

	HR (95% CI)	*P*	Allocated risk score points
Characteristics			
Del(17p)	3.8 (2.1–7.1)	<0.001	3.5
Unmutated *IGHV*	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

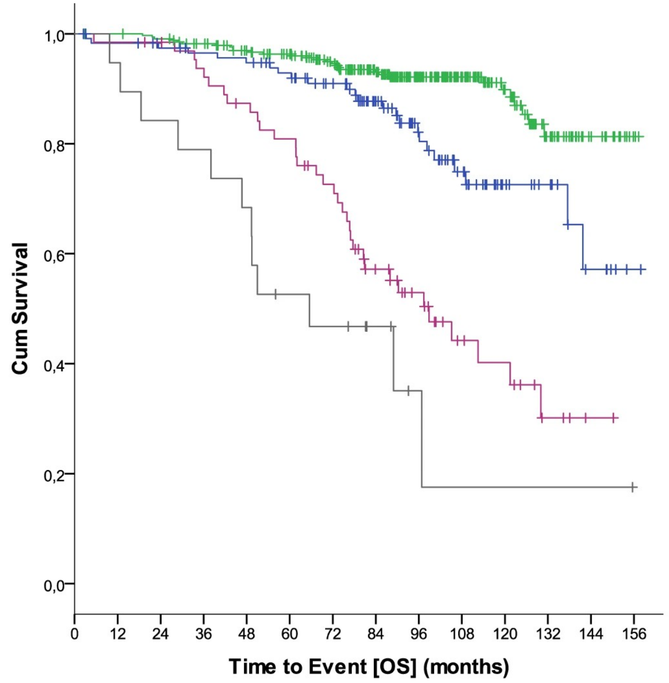
The assigned risk score points derived from the HR for OS of the individual factors.

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 accroding to the CLL1-PM		539
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		539
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
Very low	336	335	331	322	306	294	262	215	160	113	68	33	15	2
Low	119	115	111	108	106	100	89	71	49	34	19	14	6	1
High	65	64	63	59	54	50	43	29	21	12	10	4	1	0
Very high	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.

2018-12-28___Yosifov,-Stilgenbauer-et-al.__The-CLL-Success-Story__HemaSphere-Review.pdf

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. Hoechstetter[1] · Raymonde Busch[2] · Barbara Eichhorst[3] · Andreas Bühler[4] · Dirk Winkler[4] · Jasmin Bahlo[3] · Sandra Robrecht[3] · Michael J. Eckart[5] · Ursula Vehling-Kaiser[6] · Georg Jacobs[7] · Ulrich Jäger[8] · Hans Jürgen Hurtz[9] · Georg Hopfinger[10] · Frank Hartmann[11] · Harald Fuss[12] · Wolfgang Abenhardt[13] · Ilona Blau[14] · Werner Freier[15] · Lothar Müller[16] · Maria Goebeler[17] · Clemens Wendtner[1,3] · Kirsten Fischer[3] · Carmen D. Herling[3] · Michael Starck[1] · Martin Bentz[18] · Bertold Emmerich[19] · Hartmut Döhner[20] · Stephan Stilgenbauer[20] · Michael Hallek[3]

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		*P* value
			Lower	Upper	
COX regression OS					
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
COX regression TTFT					
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.**	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.

	HR (95% CI)	*P*	Allocated risk score points
Characteristics			
Del(17p)	3.8 (2.1–7.1)	<0.001	3.5
[illegible]	[illegible]	[illegible]	[illegible]
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

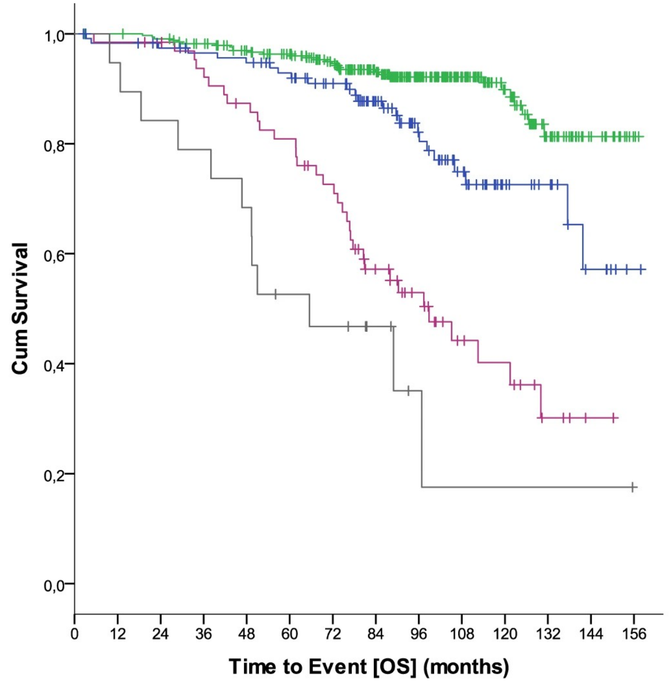
The assigned risk score points derived from the HR for OS of the individual factors.

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 accroding to the CLL1-PM		539
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		539
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
Very low	336	335	331	322	306	294	262	215	160	113	68	33	15	2
Low	119	115	111	108	106	100	89	71	49	34	19	14	6	1
High	65	64	63	59	54	50	43	29	21	12	10	4	1	0
Very high	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.

2018-12-28___Yosifov,-Stilgenbauer-et-al.__The-CLL-Success-Story__HemaSphere-Review.pdf

Cancer Classifications & Parameters

NCIt | ICD-O / WHO | TNM

International Agency for Research on Cancer
World Health Organization

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

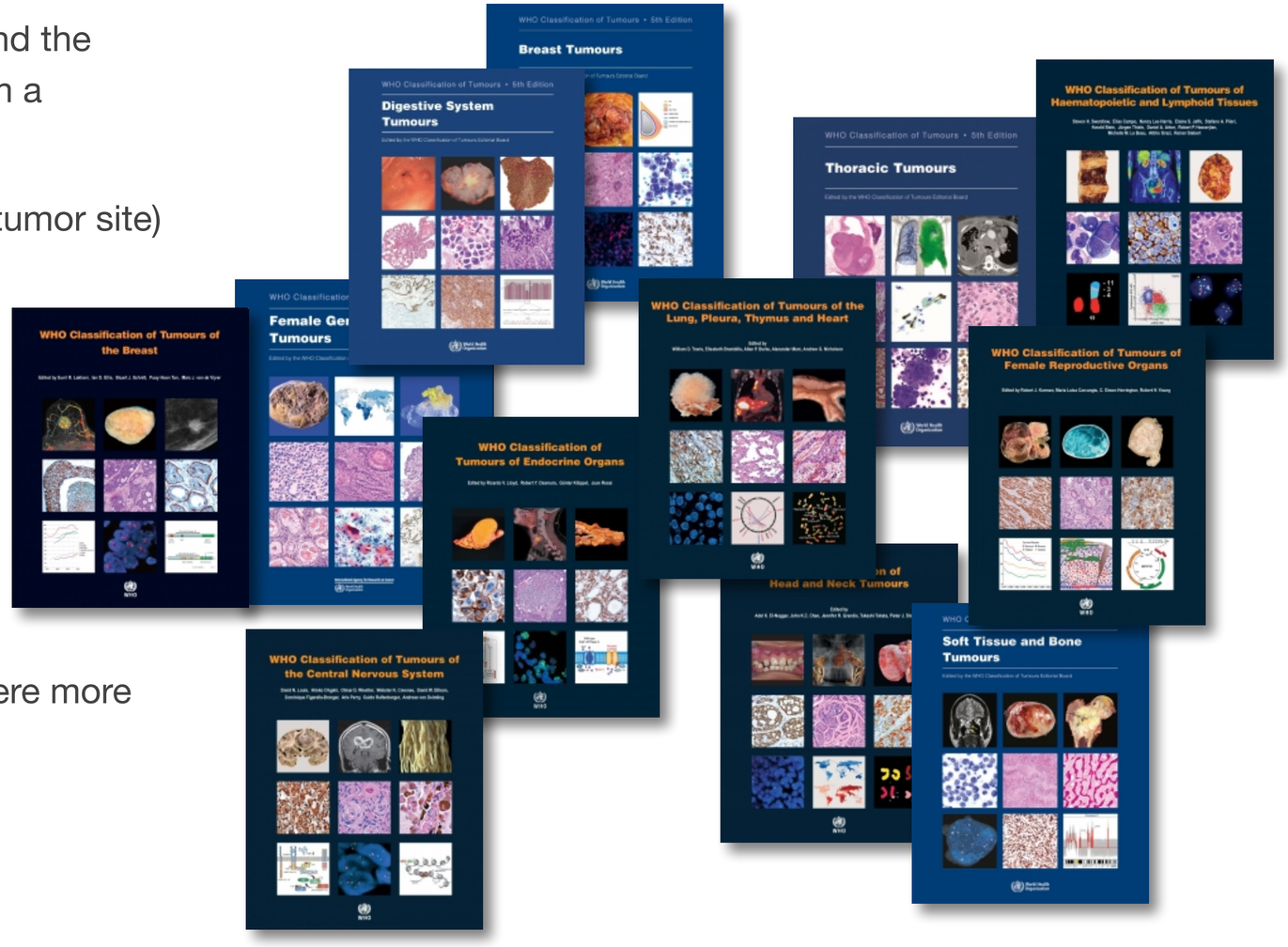
  - "Adenocarcinoma" of the "Sigmoid colon"

    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



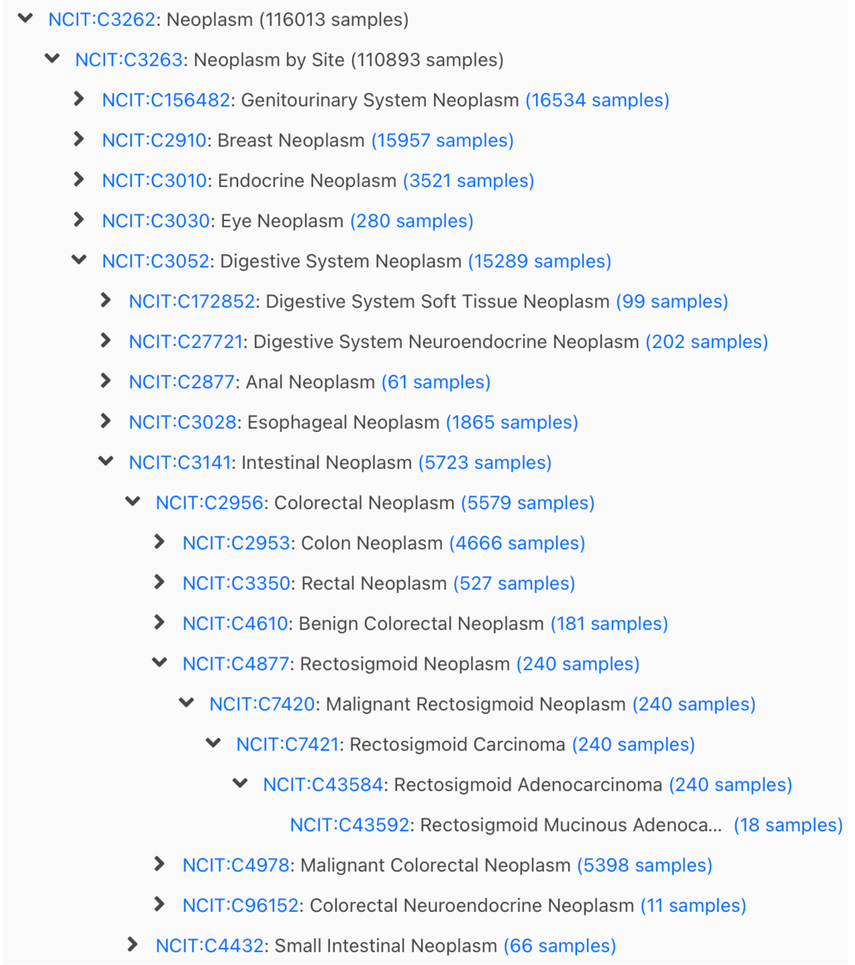
NCIt

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

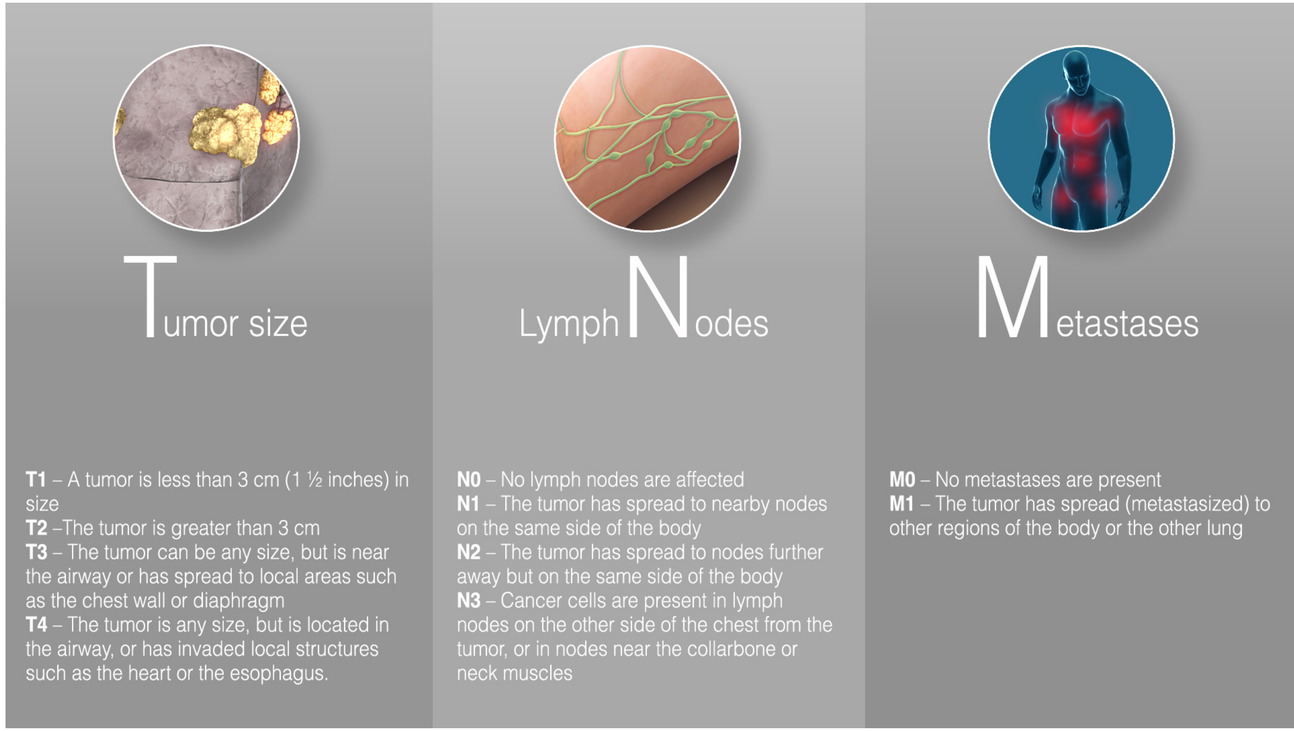
  - **NCIT:C43584** - Rectosigmoid Adenocarcinoma
  - **NCIT:C7541** - Retinoblastoma
- truly hierarchical ontology
- hierarchical system empowers "logical OR" queries
- terms can have multiple occurrences in diagnostic tree
- assignment of code to different groupings allows soft aggregation (e.g. a type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



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

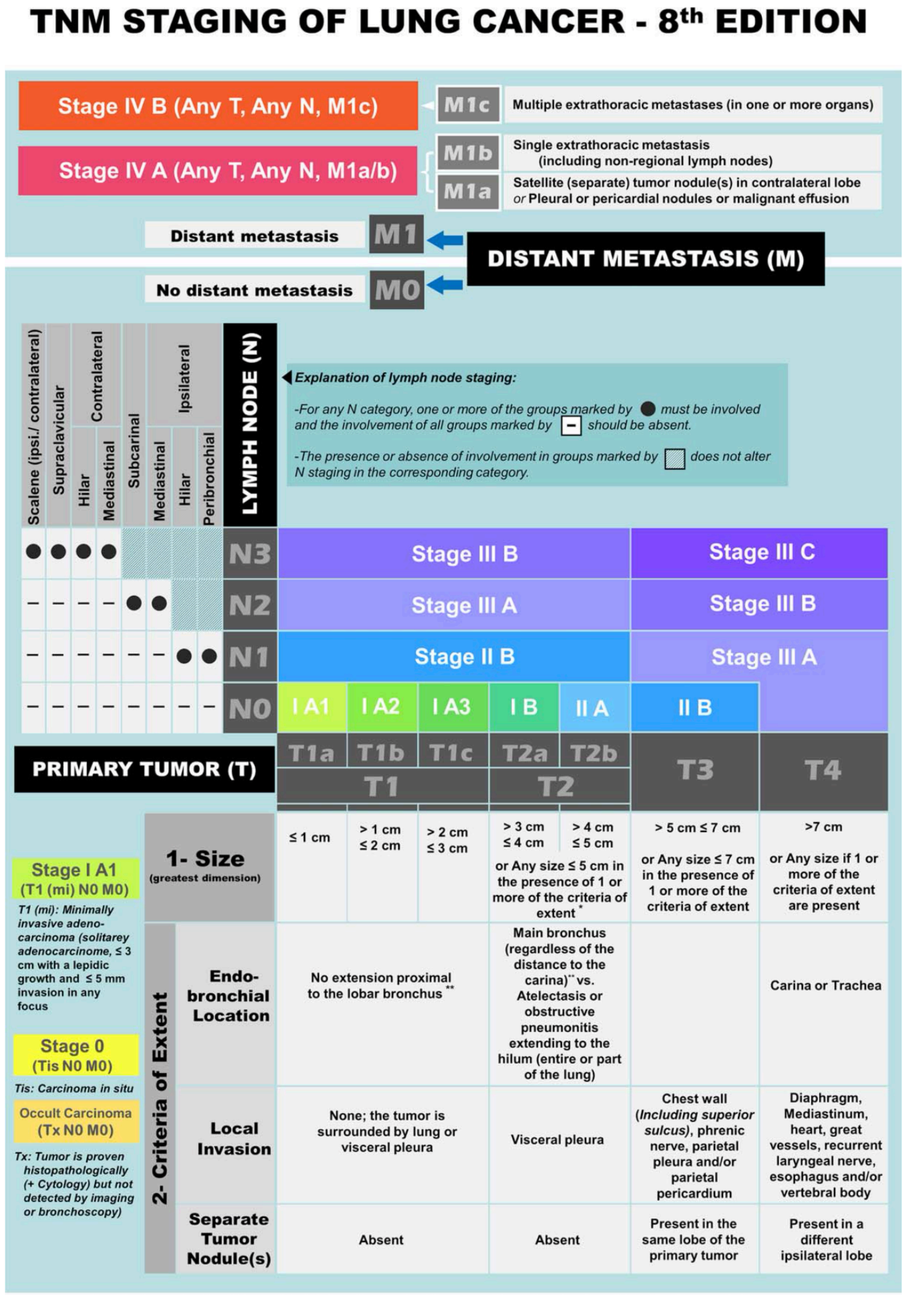
Source: https://www.cancer.gov/about-cancer/diagnosis-staging/staging

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

Source: https://www.cancer.gov/about-cancer/diagnosis-staging/staging



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

Source: https://www.cancer.gov/about-cancer/diagnosis-staging/staging

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:C48699	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	N1bI Stage Finding	2
NCIT:C95935	N1bII Stage Finding	3
NCIT:C95936	N1bIII 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	

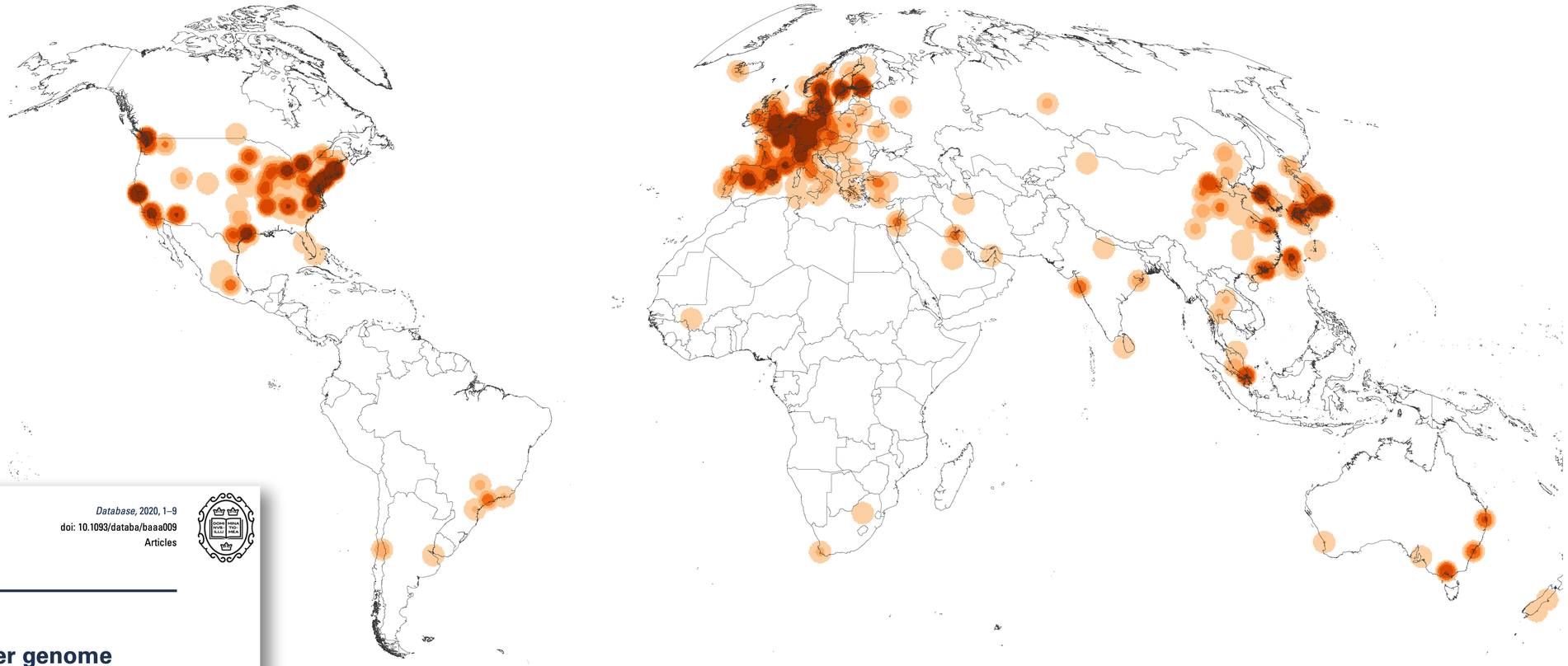
Tasks

Survival analyses | Cancer classifications | Staging

- Familiarize yourself with the different concepts behind different disease clasification systems - what are there use, advantages, problems? E.g. ICD-10, ICD-O, NCIt
  - 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?



DATABASE
The Journal of Biological Databases and Curation

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

Articles

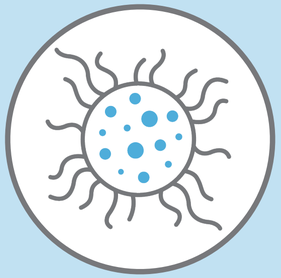
Geographic assessment of cancer genome profiling studies

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 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

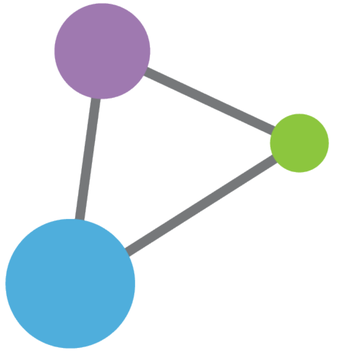
Advance precision medicine

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

progenetix

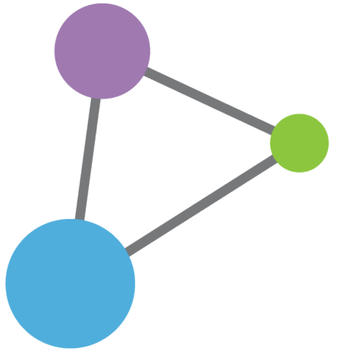
European Genome-Phenome Archive

SPHN Swiss Personalized Health Network

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

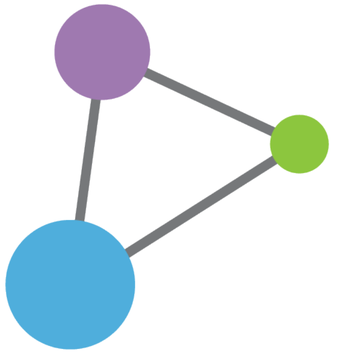
Centralized Genomic Knowledge Bases

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.net: storage and visualization of genomic aberration data in human malignancies

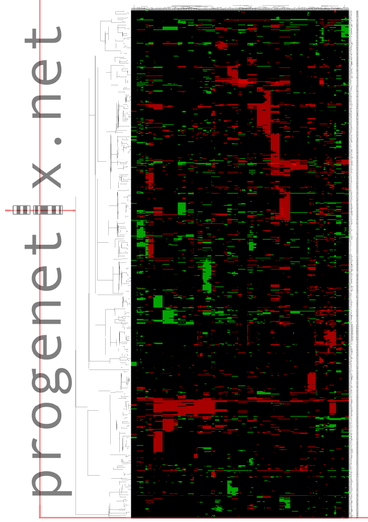
michael baudis, md

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).



Data selection
PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data
Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage
Currently, the primary data is stored in a dedicated "off-line" database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

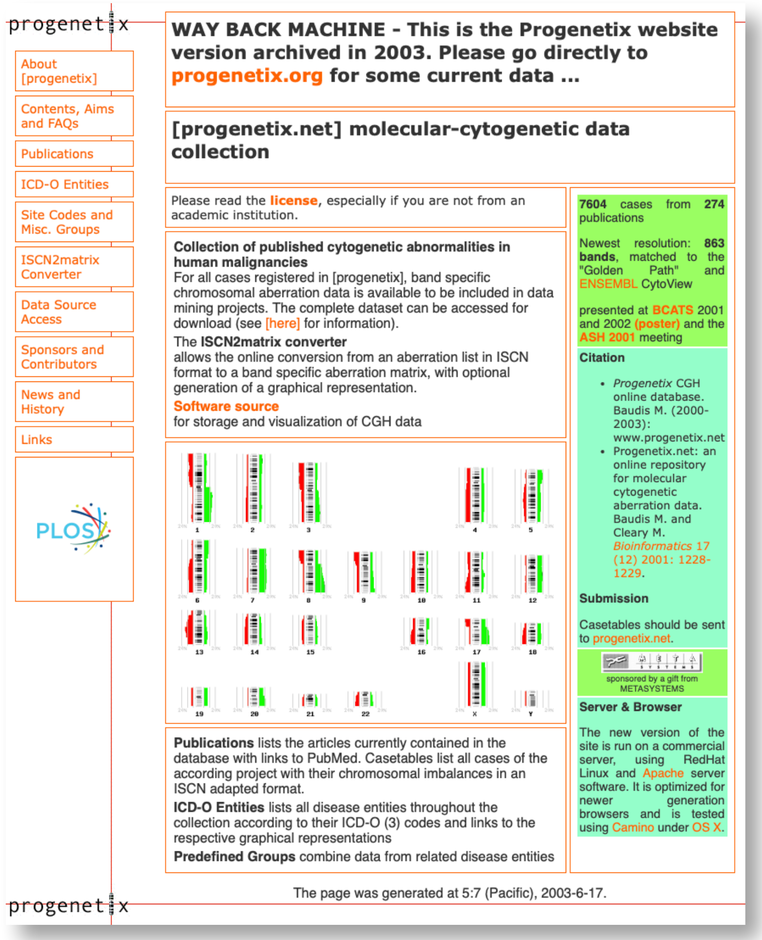
Text parsing and generation of aberration matrix
For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with "1" and losses with "–1"; localized high-level gains are designated "2".

Website generation
For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.

Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

Michael Baudis
BCATS Biocomputing at Stanford
Stanford Nov 2001

progenetix.net website from June 2003



BIOINFORMATICS **APPLICATIONS NOTE** Vol. 17 no. 12 2001
Pages 1228–1229

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

Michael Baudis[1,2,*] *and Michael L. Cleary*[2]

[1]*Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and* [2]*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 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; Kallioniemi *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 comparision 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[†].

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

*To whom correspondence should be addressed.

[†]Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.


```
Domain Name: PROGENETIX.NET
Registry Domain ID: 45628826_DOMAIN_NET-VRSN
Registrar WHOIS Server: whois.enterprice.net
Registrar URL: http://www.epag.de
Updated Date: 2023-11-30T08:32:22Z
Creation Date: 2000-11-29T18:17:38Z
```

progenetix.org

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. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics

progenetix

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

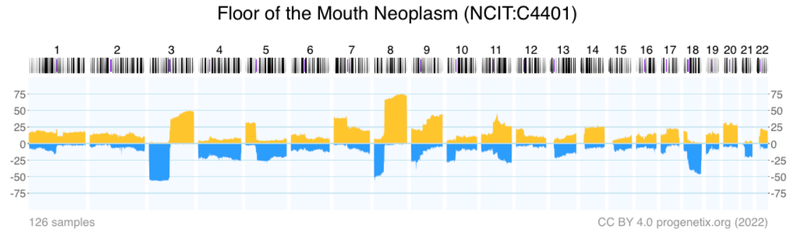
Beacon+

Documentation
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

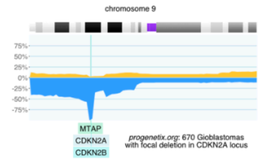
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.



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.

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
```

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

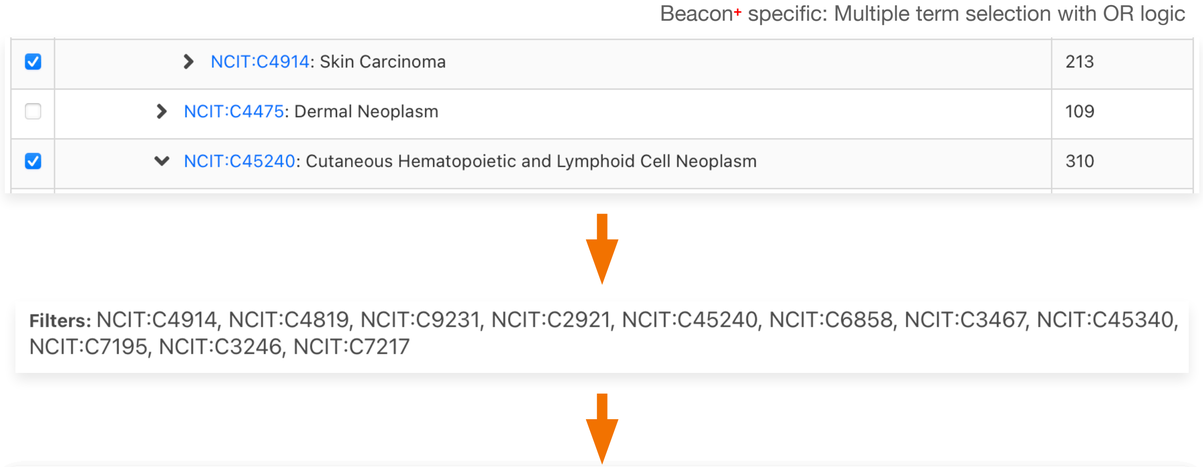
Beacon v2 Filters

Example: Use of hierarchical classification systems (here NCIt 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

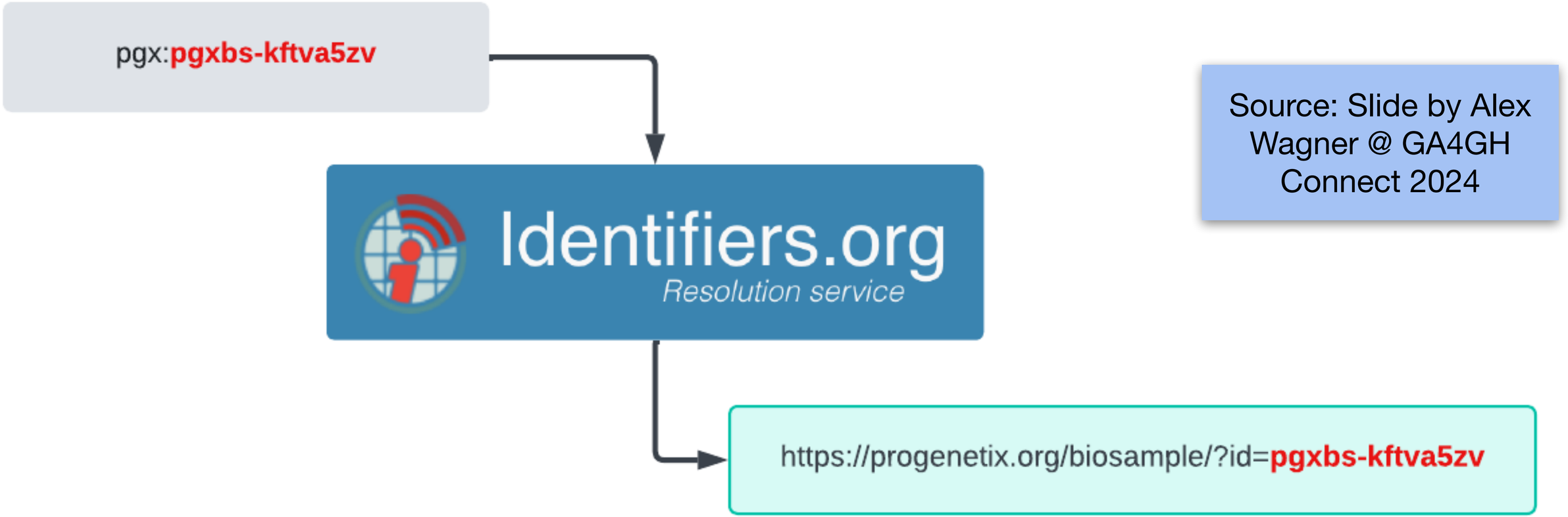
  ➡ implicit *OR* with otherwise assumed *AND*
- implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations

Beacon+ specific: Multiple term selection with OR logic



Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

CURIE Resolution: Identifiers.org



Source: Slide by Alex Wagner @ GA4GH Connect 2024

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

progenetix

Different Approaches to Data Sharing

progenetix

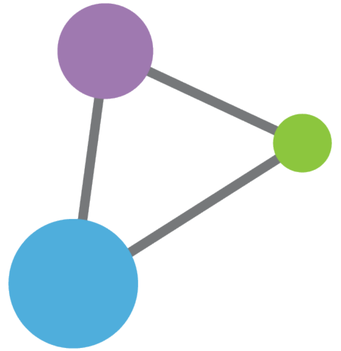
Centralized Genomic Knowledge Bases

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 EGA

European Genome-Phenome Archive

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*")

EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

EMBL-EBI

Global Alliance for Genomics & Health

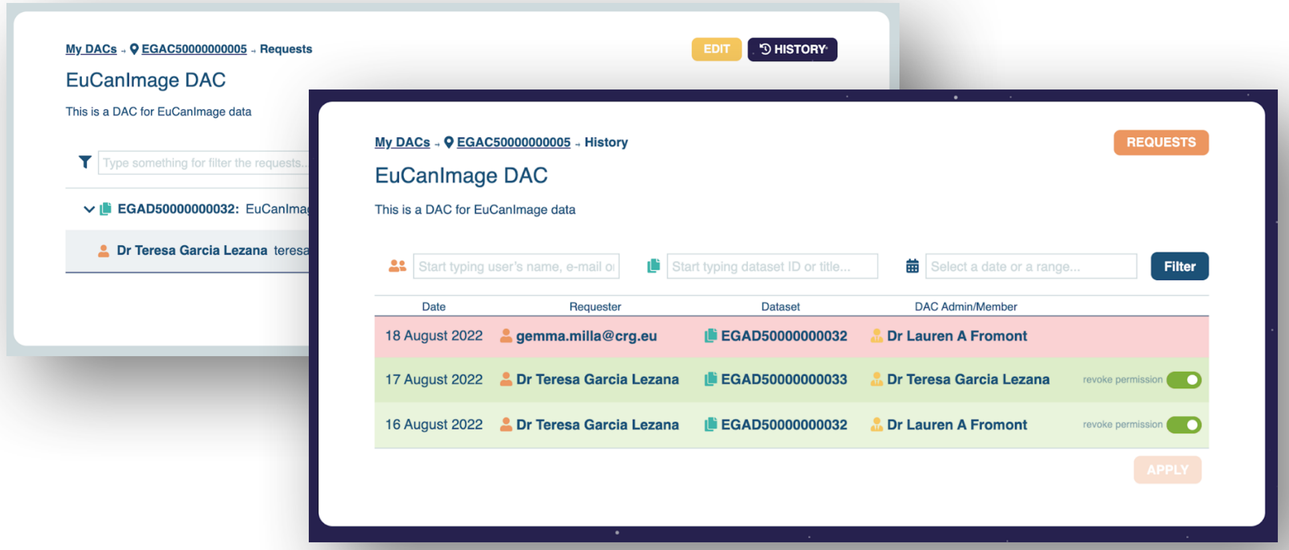
DRIVER PROJECT

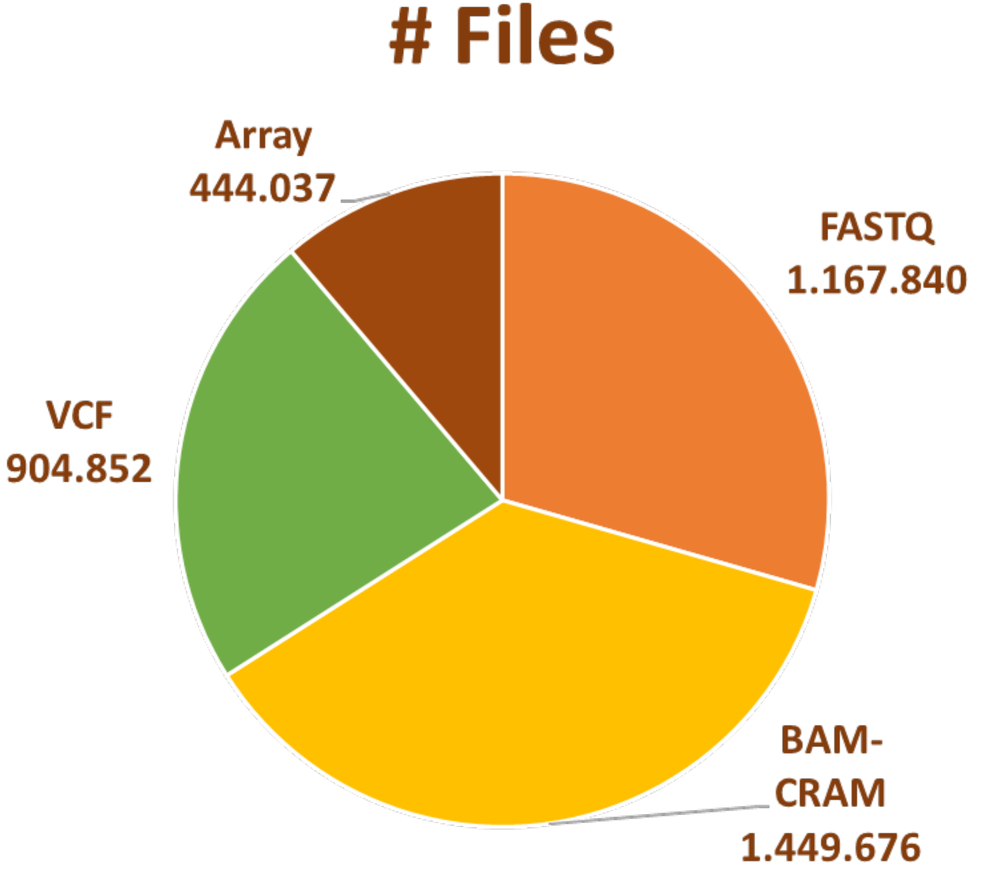
Slide: adapted from Jordi Rambla@ GA4GH 2023

The EGA

European Genome-Phenome Archive

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





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

Slide: adapted from Jordi Rambla@ GA4GH 2023

Different Approaches to Data Sharing

progenetix

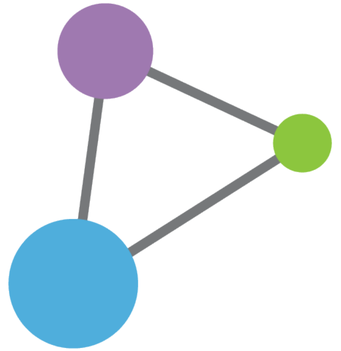
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

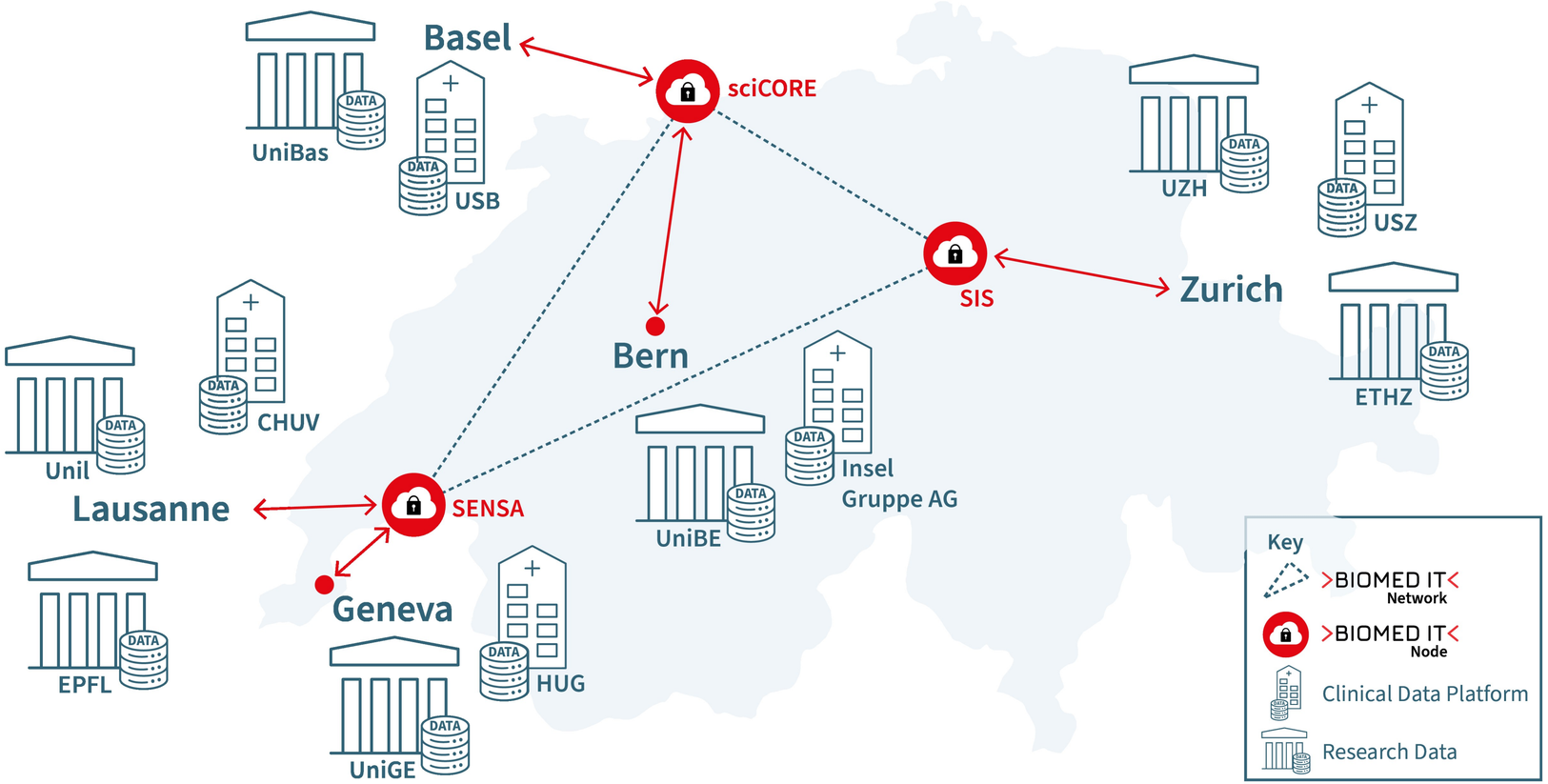
SPHN Swiss Personalized Health Network

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



Linkage of distributed and disparate datasets

The Swiss Personalized Health Network



Strategic Focus Area
Personalized Health and Related Technologies

Personalized Health Informatics Group

SPHN Data Coordination Center (DCC)
BioMedIT Network

Different Approaches to Data Sharing

progenetix

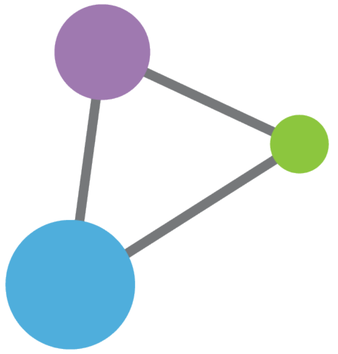
European Genome-Phenome Archive

SPHN Swiss Personalized Health Network

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

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

INFORMATICS

Beacon v2 and Beacon networks: federated data discovery n biome

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

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

OP

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,*]

A New Paradigm for Data Sharing

FROM

TO

Data Copying

Data Visiting

A New Paradigm for Data Sharing

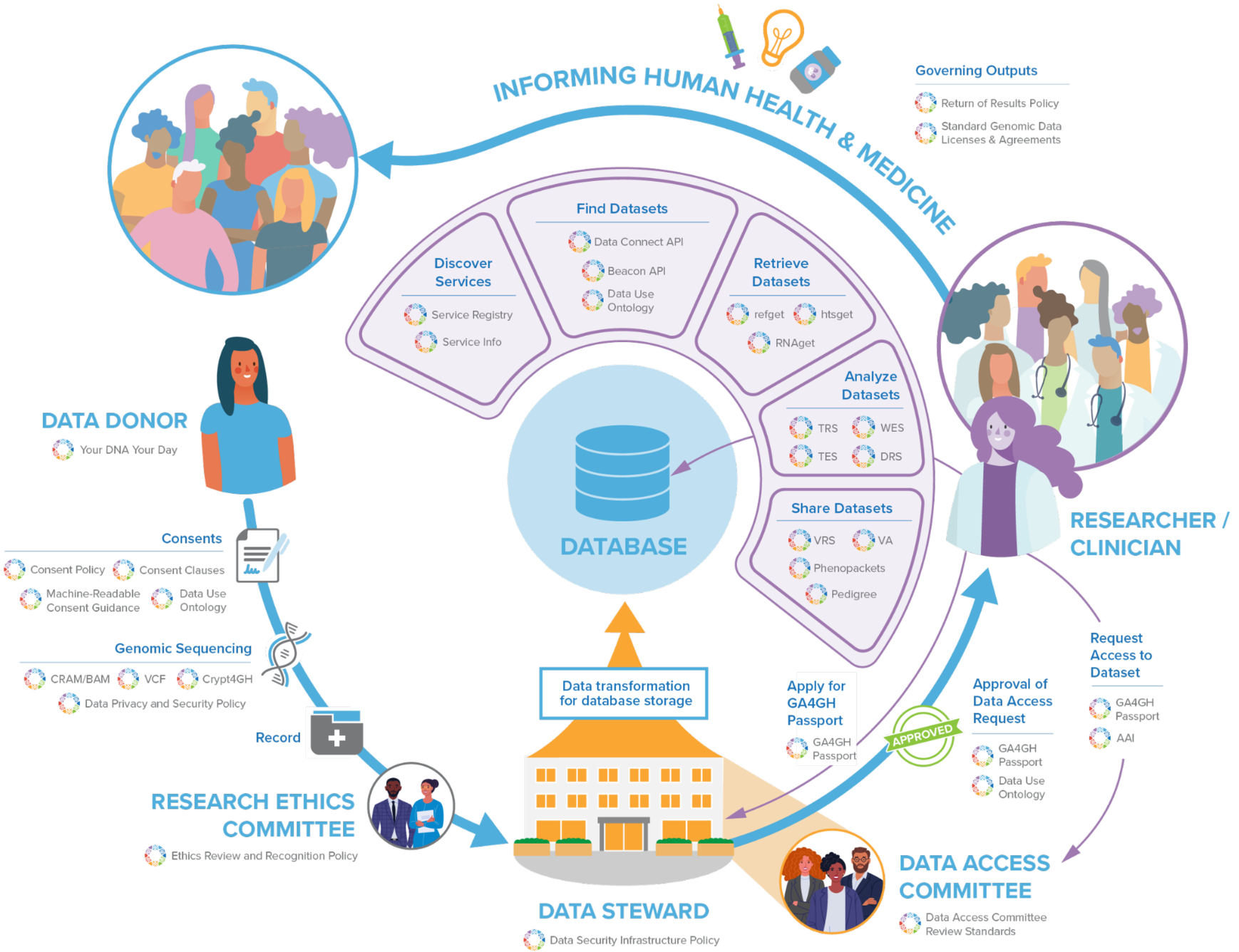
FROM

STANDARDS

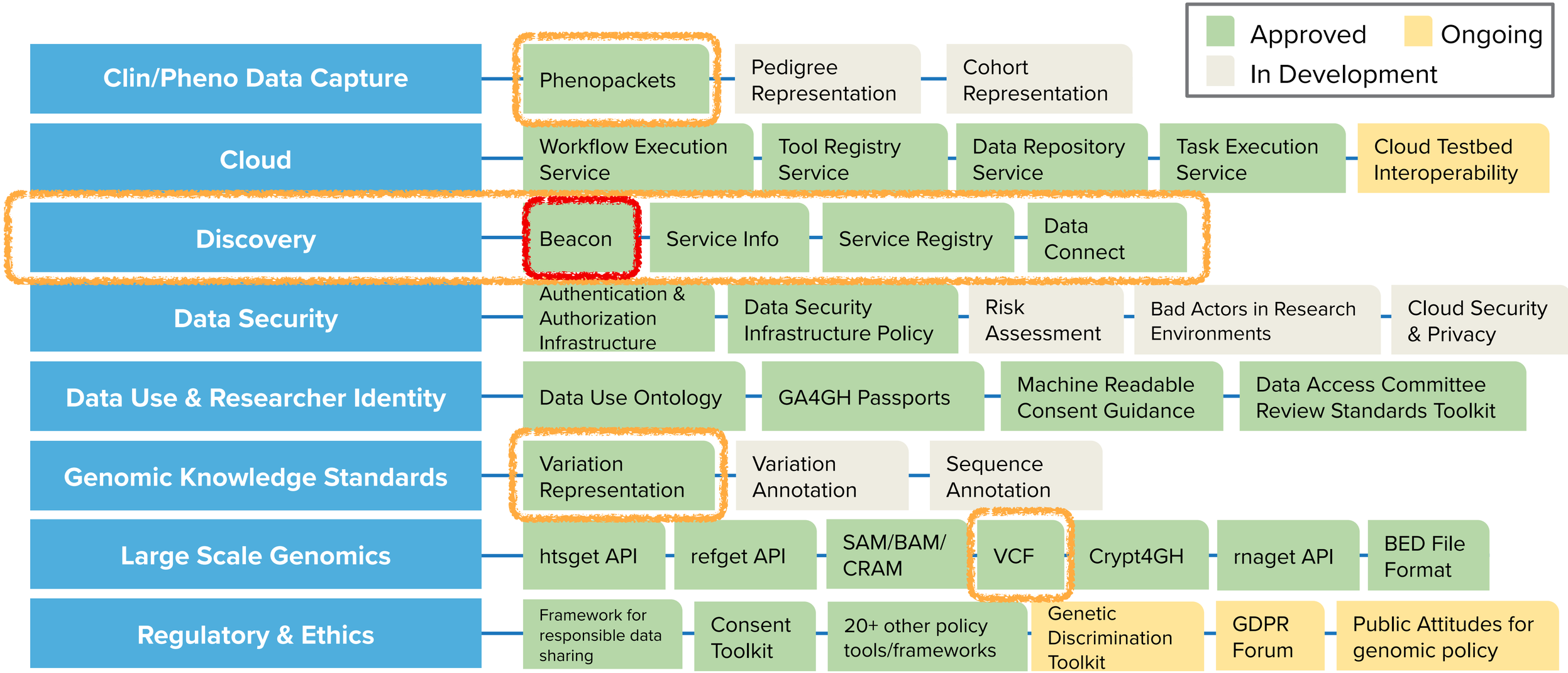
TO

Data Copying

Data Visiting


INFORMING HUMAN HEALTH & MEDICINE
Governing Outputs
Return of Results Policy
Standard Genomic Data Licenses & Agreements
Find Datasets
Data Connect API
Beacon API
Data Use Ontology
Discover Services
Service Registry
Service Info
Retrieve Datasets
refget
htsget
RNAget
Analyze Datasets
TRS
WES
TES
DRS
Share Datasets
VRS
VA
Phenopackets
Pedigree
DATABASE
DATA DONOR
Your DNA Your Day
RESEARCHER / CLINICIAN
Consents
Consent Policy
Consent Clauses
Machine-Readable Consent Guidance
Data Use Ontology
Genomic Sequencing
CRAM/BAM
VCF
Crypt4GH
Data Privacy and Security Policy
Record
Data transformation for database storage
Apply for GA4GH Passport
GA4GH Passport
APPROVED
Approval of Data Access Request
GA4GH Passport
Data Use Ontology
Request Access to Dataset
GA4GH Passport
AAI
RESEARCH ETHICS COMMITTEE
Ethics Review and Recognition Policy
DATA STEWARD
Data Security Infrastructure Policy
DATA ACCESS COMMITTEE
Data Access Committee Review Standards

Overview of GA4GH standards and frameworks



VCF/BCF

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. The Binary Call Format (BCF) is the Binary equivalent, smaller and more efficient to process.

Software Libraries: htslib | htsjdk

Tools: Samtools | BCFtools

Databases: European Variation Archive (EVA) | dbGAP | dbSNP | 1000 Genomes Projects / IGSR

Genome Browsers: ENSEMBL | JBrowse | UCSC Genome Browser

Example Users

All of Us RESEARCH PROGRAM

BROAD INSTITUTE

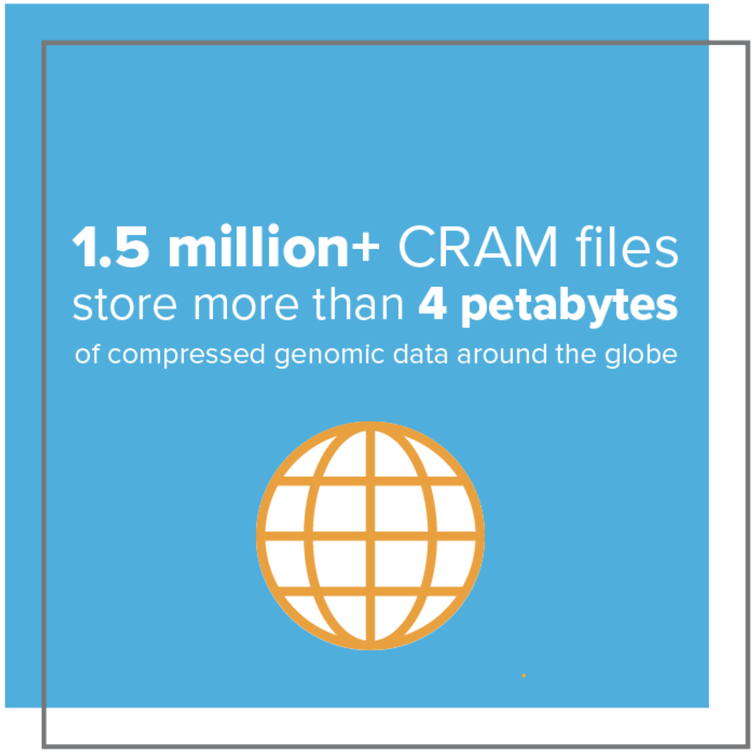
elixir

EMBL-EBI

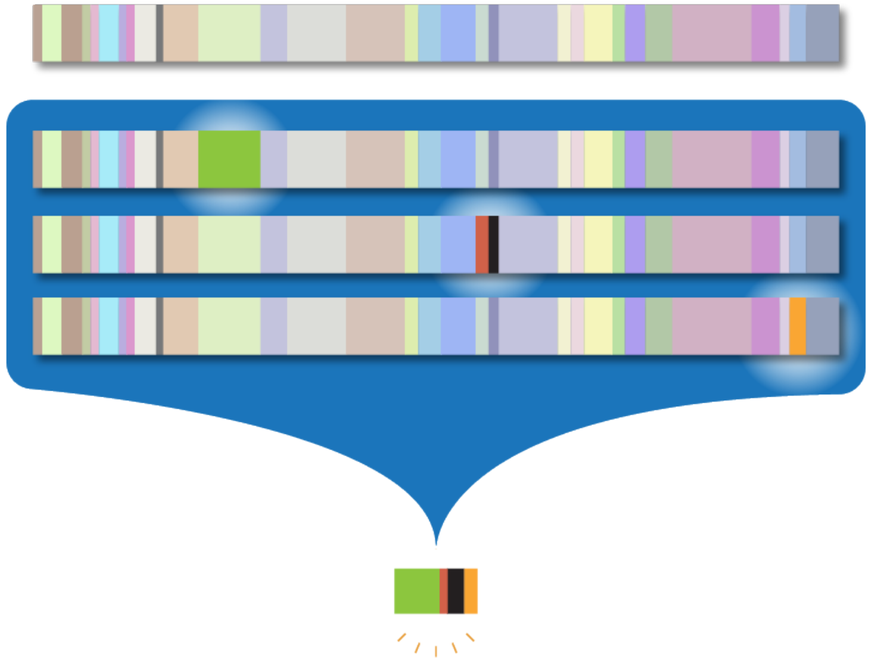
Genomics england

CRAM

CRAM is a file format for storing compressed genomic data. To make files small and efficient, the algorithm compresses information by only storing the parts that are different from the reference human genome.



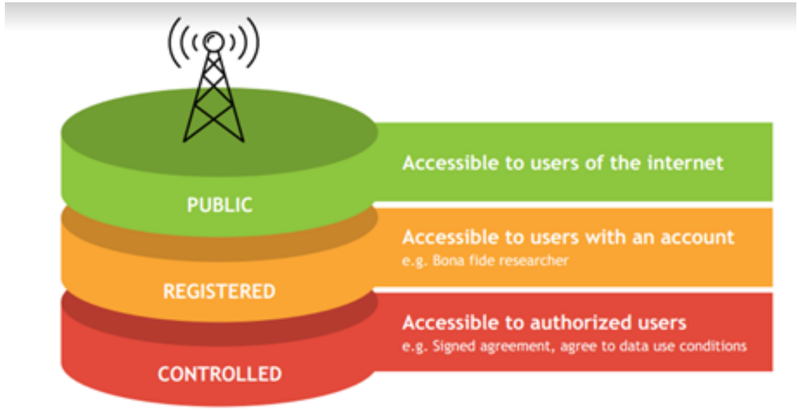
CRAM compresses data by only storing the difference.

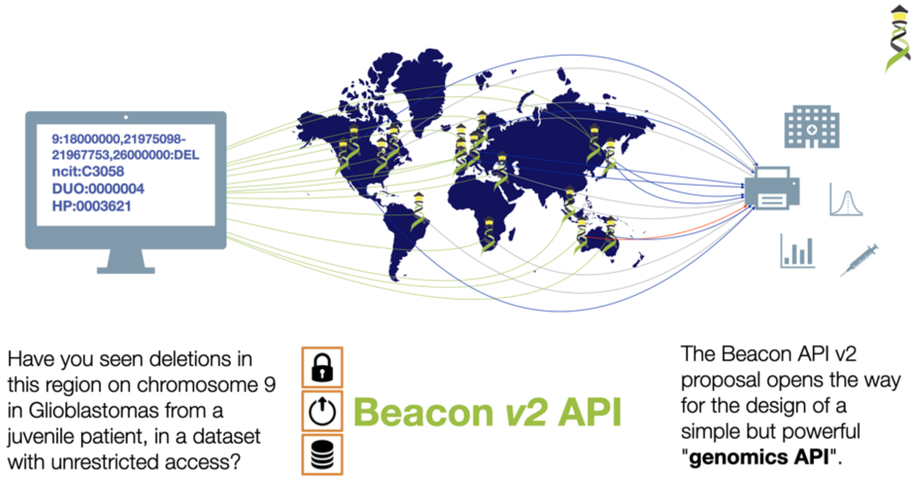


Beacon API v2

The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

Approved: April 21, 2022





Example Users

EMBL-EBI

Australian Genomics

SciLifeLab

elixir

UNIVERSITY OF CALIFORNIA SANTA CRUZ

BROAD INSTITUTE

EUROPEAN GENOME-PHENOME ARCHIVE

International Cancer Genome Consortium

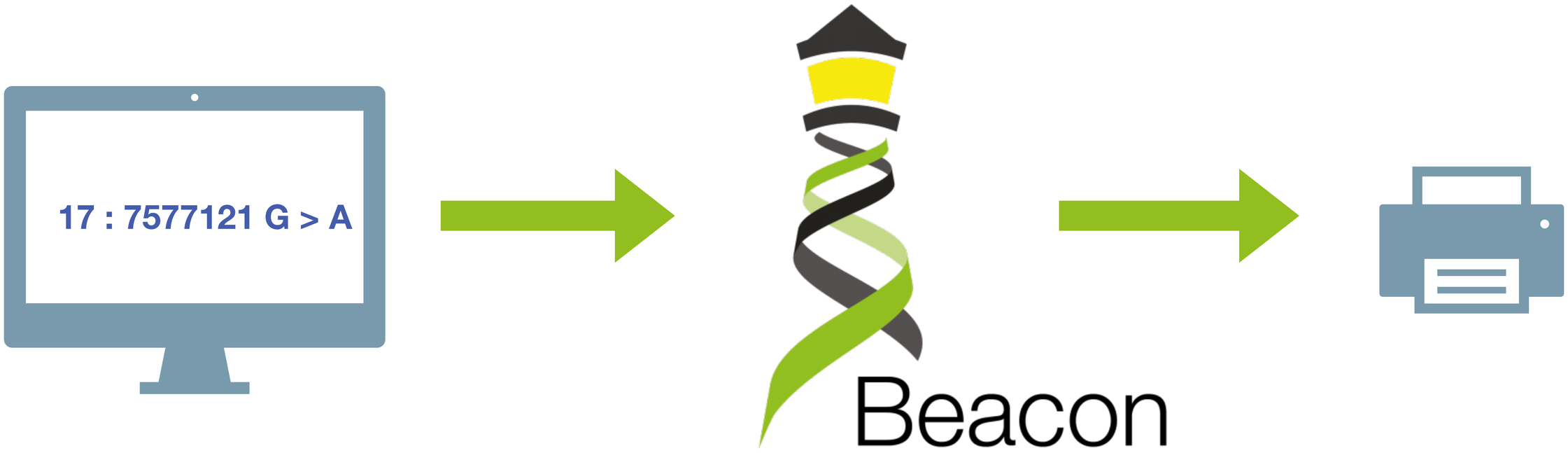
Beacon v2 and Beyond

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

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

elixir

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	**GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"**		
2015	• beacon-network.org aggregator created by DNAstack		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented @ **progenetix.org** • 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 • adopting "handover" concept	• new Beacon website (March)
2019	• ELIXIR Beacon Network		• Beacon publication at Nature Biotechnology
2020		• "**Scouts**" teams working on different aspects - filters, genomic variants, compliance ...	
2021		• **framework** + **models** concept implemented • range and bracket queries, variant length • starting of GA4GH review process	• Phenopackets v2 approved
2022		• changes in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved** at April GA4GH Connect	• *docs.genomebeacons.org*



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

YES | **NO** | **\0**

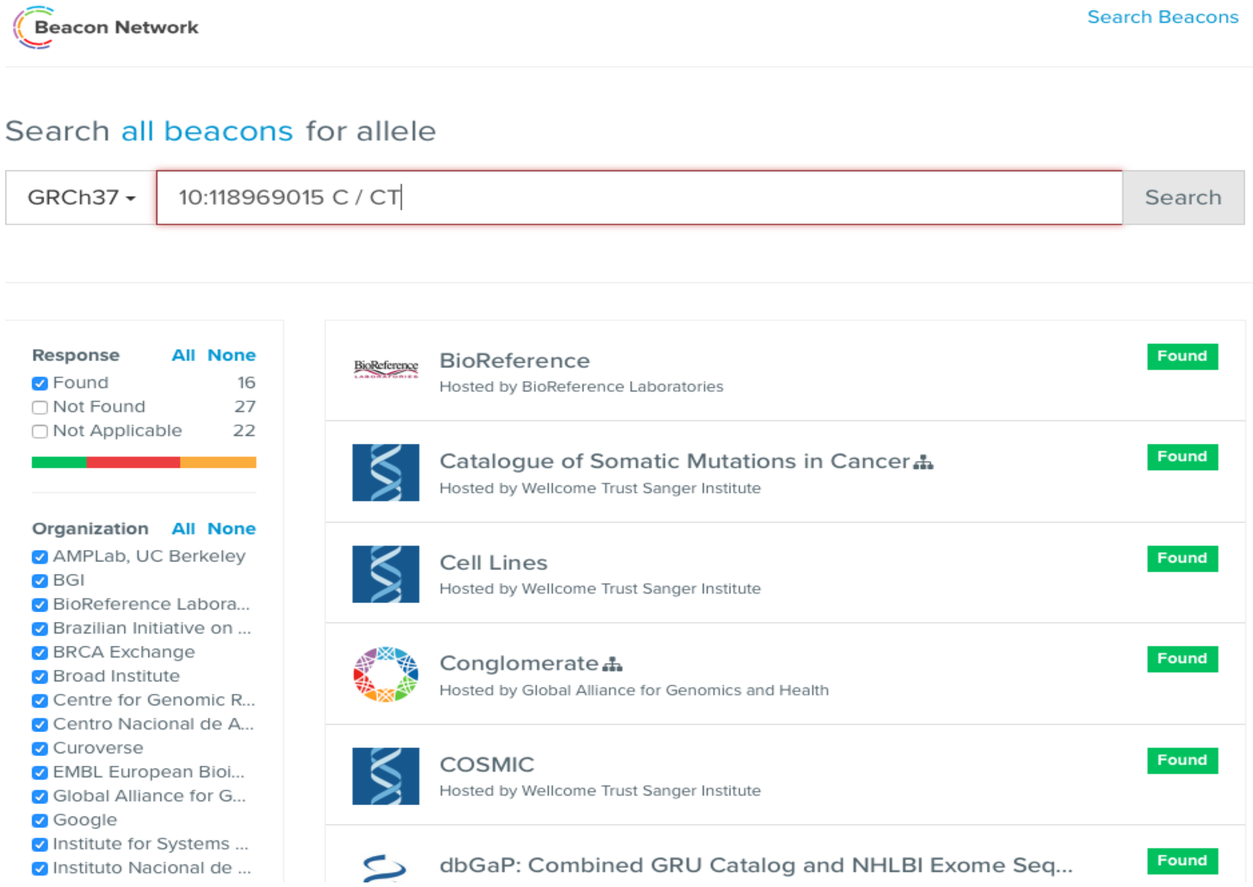
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

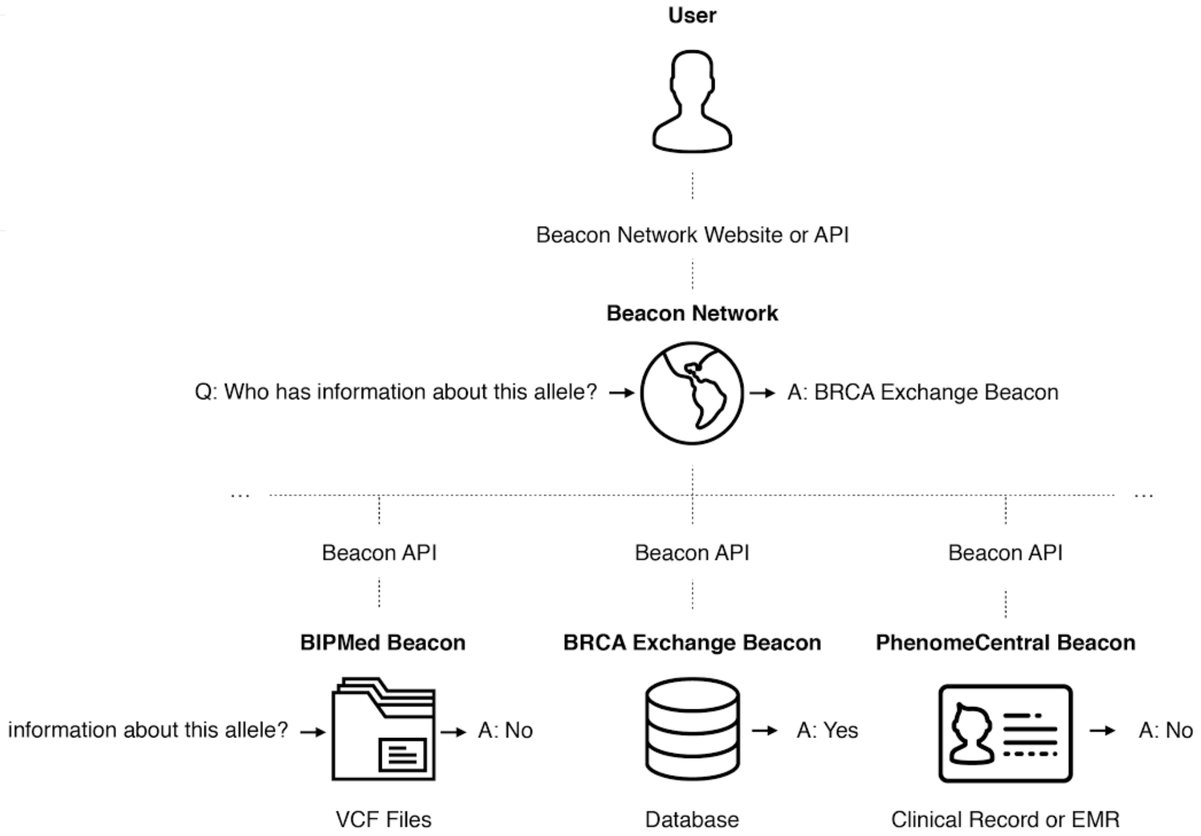
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

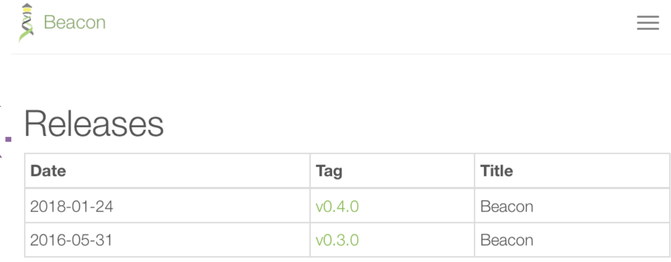
Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon Project in 2016

An open web service that tests the willingness of international sites to share genetic data.







35+ Organizations

90+ Beacons

200+ Datasets

100K+

Global Alliance for Genomics & Health

Global Alliance “Beacon” - Jim Ostell, NCBI, March 7, 2014

Global Alliance for Genomics & Health

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.

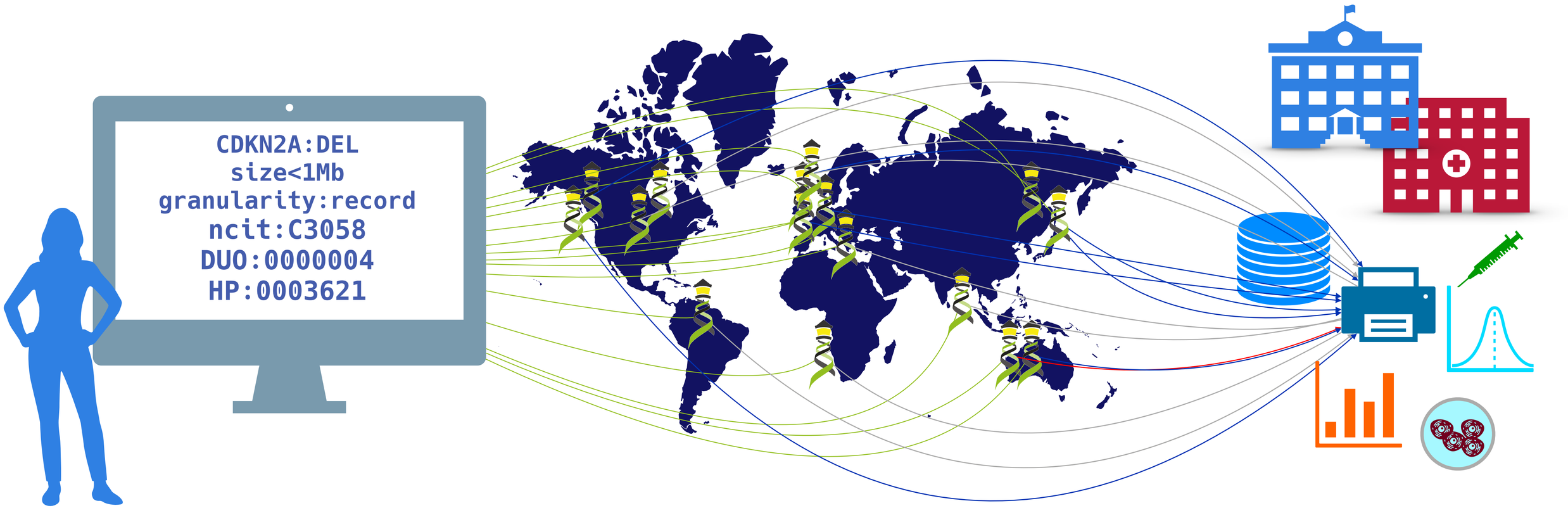
A number of more useful first versions have been suggested.

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*** …



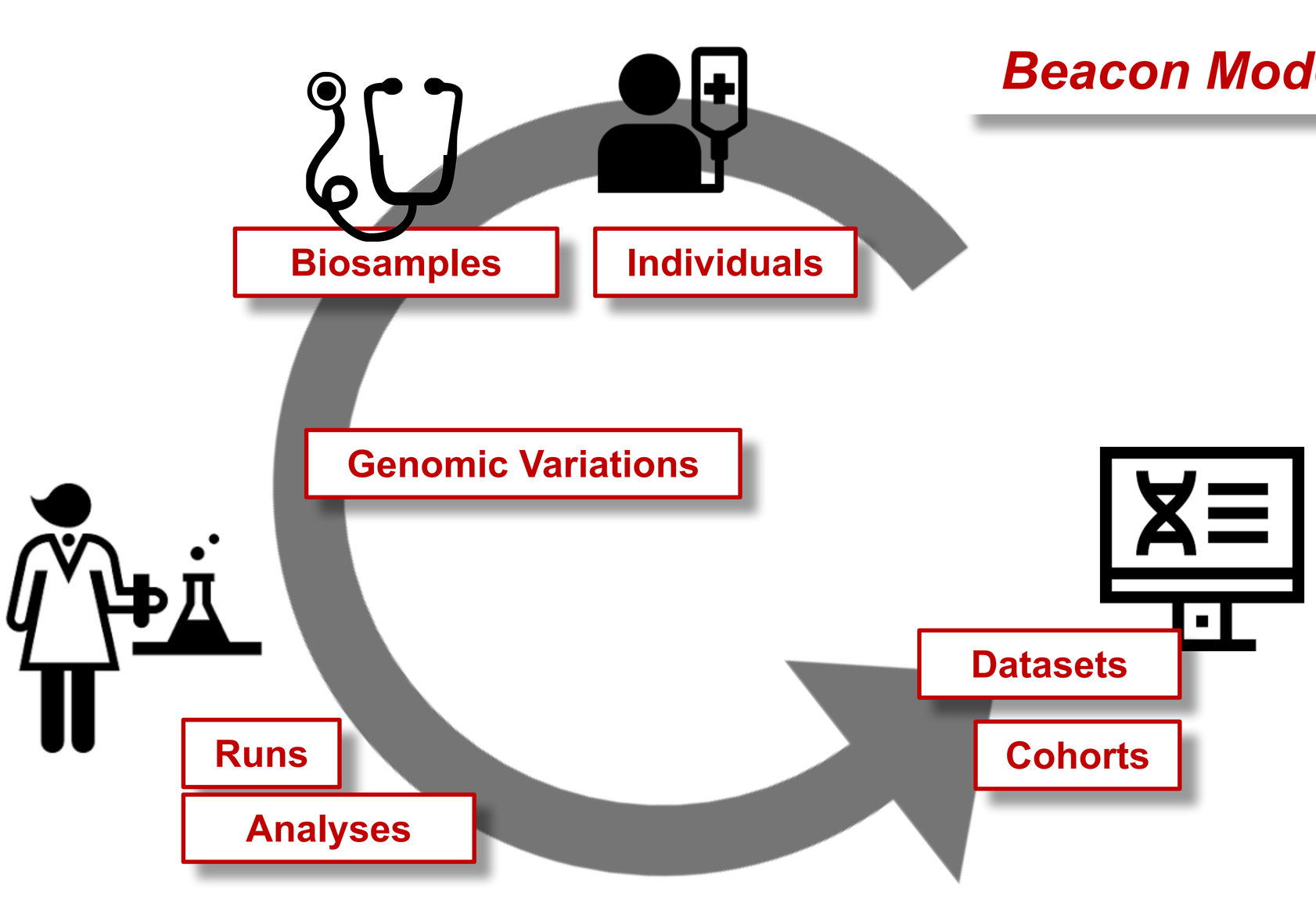
Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

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 Model

Beacon Framework (protocol)

Begriffsbestimmung

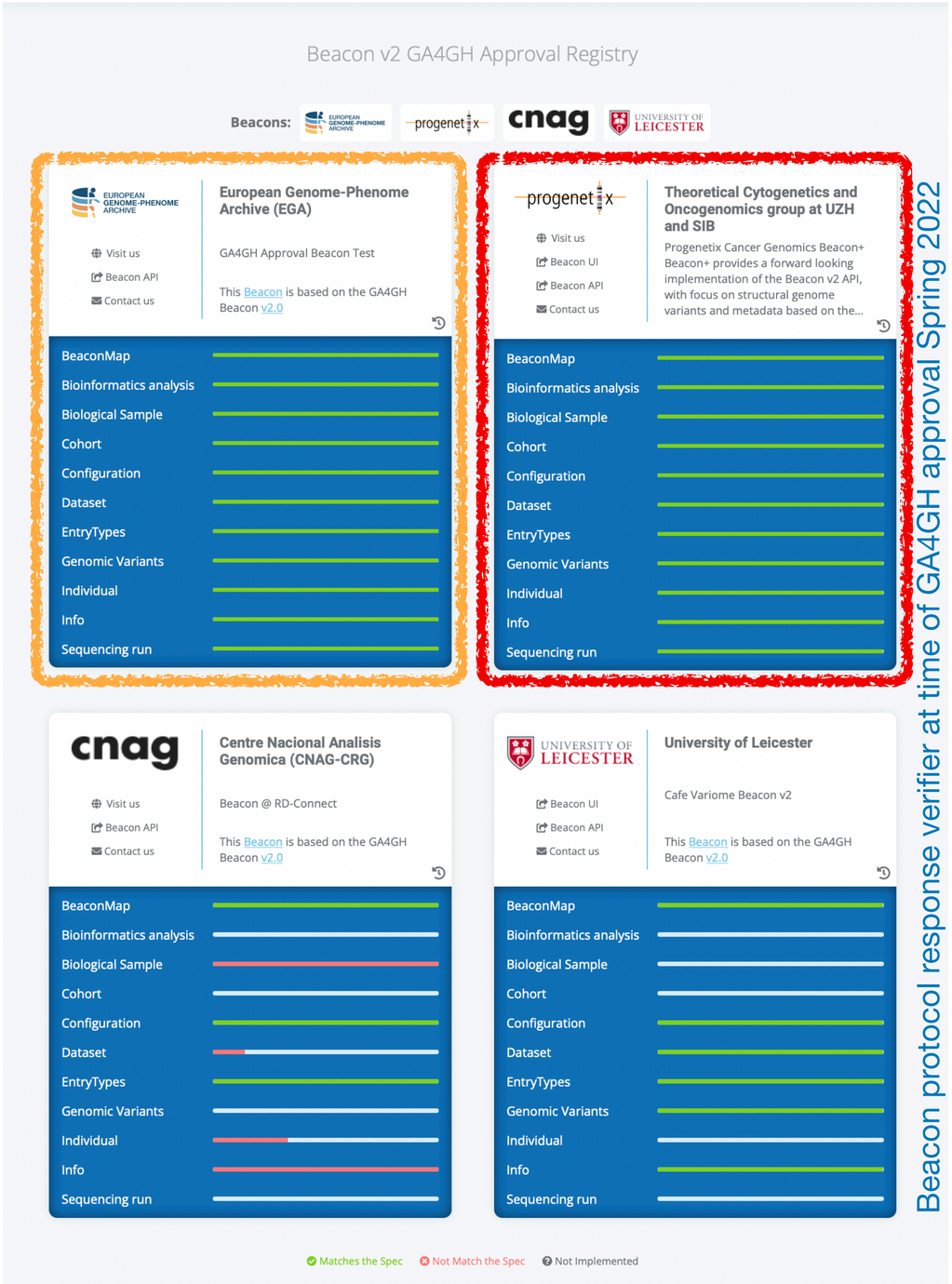
The right expressions help to conceptualize...

- **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 wit 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
  - ➡ ...



Global Alliance for Genomics & Health

progenetix

elixir

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 autofills
- *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

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

Entity collections

Utility collections

github.com/progenetix/bycon GitHub

pgxRpi: an R/Bioconductor package

Client for Accessing Beaconized Data

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

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>

- **Query and export 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**

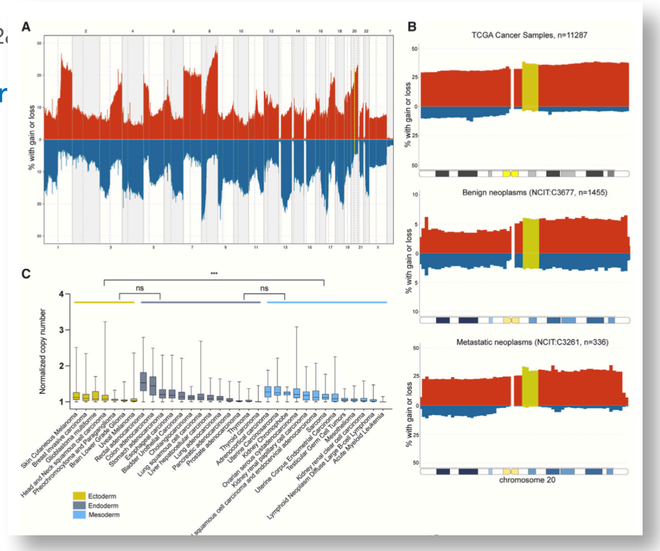
http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C3512

```
> 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)
```

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

ISSCR

OPEN ACCESS

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

Nuša Krivec,[1,2] Manjusha S. Ghosh,[1,2] and Claudia Spits[1,2,*]

[1]Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

[2]These authors contributed equally

*Correspondence: 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

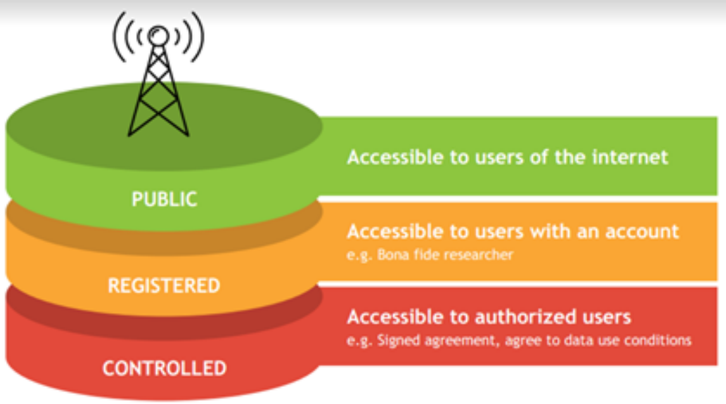
Making Beacons Biomedical - Beacon v2

- 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

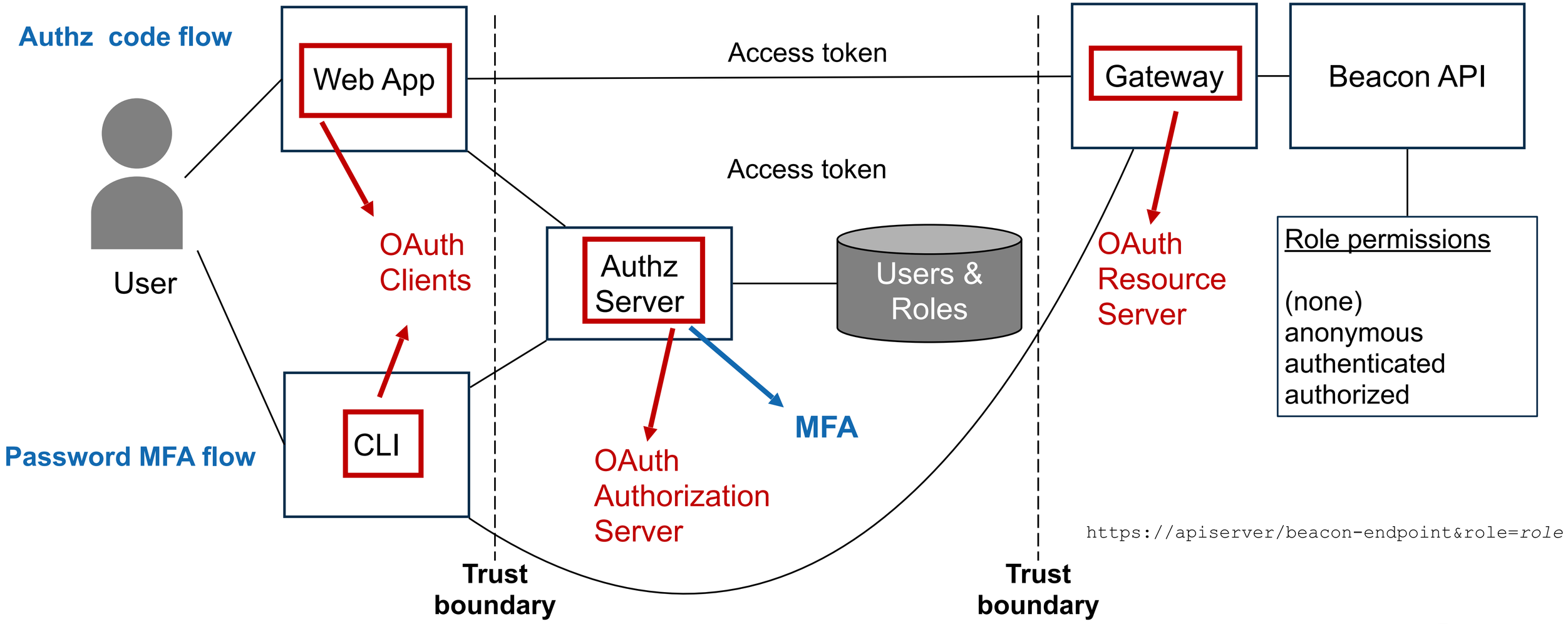
Security by Design ... if Implemented in the Environment



- 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 autorization 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