Connecting the silos

Genomic Data Standards, Resources and the Global Alliance for Genomics and Health

University of Zurich[UZH]

Global Alliance for Genomics & Health

1992 — Heidelberg

Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics

2001 — Stanford

Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website

2003 — Gainesville

Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis

2006 — Aachen

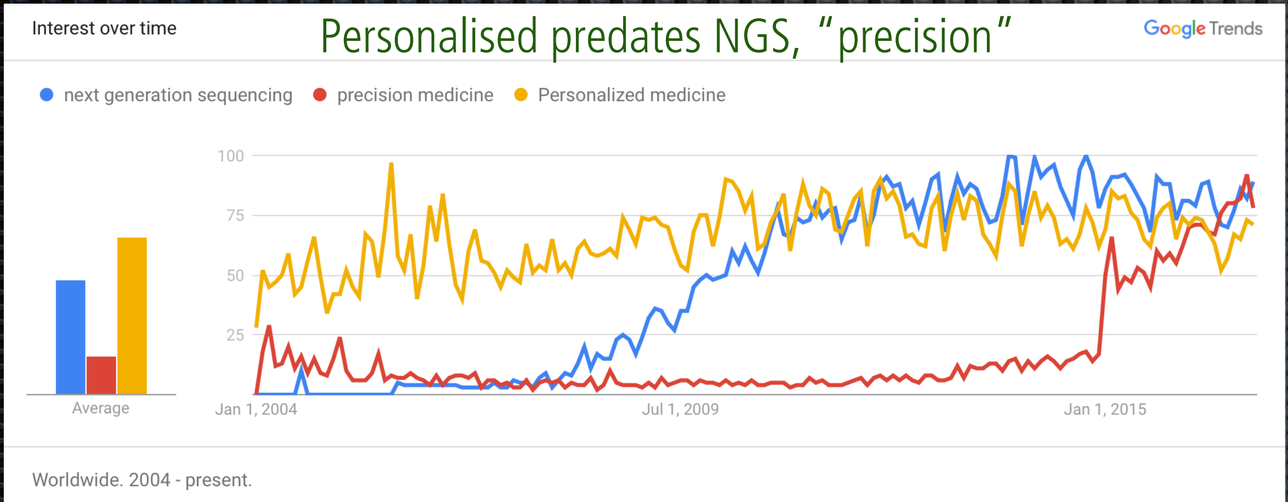
Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities

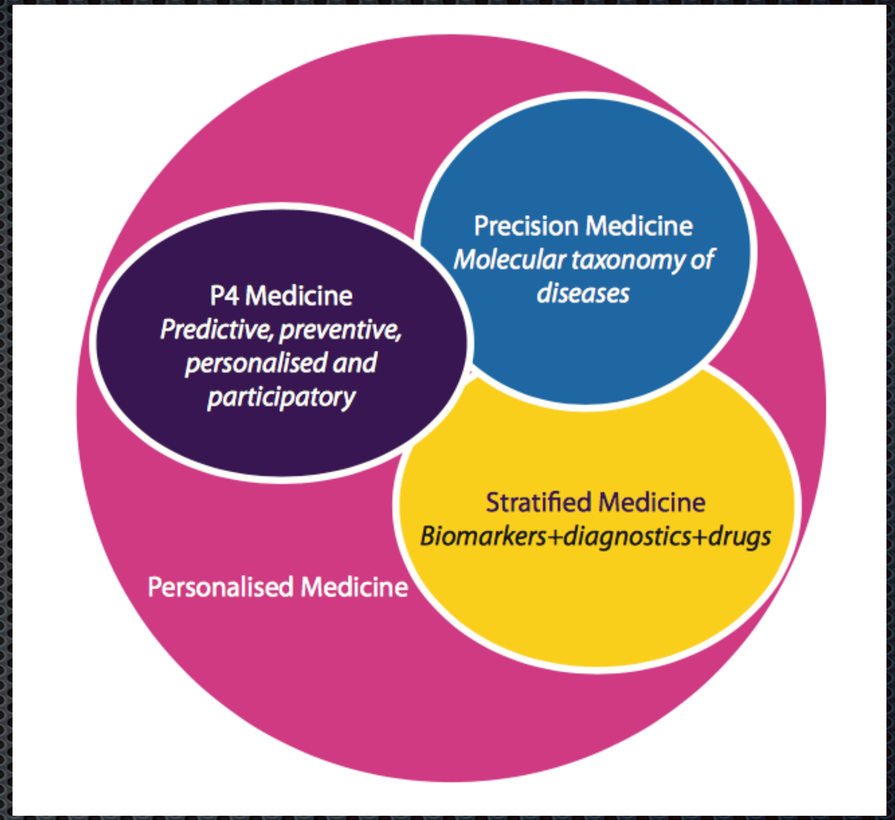
2007 — Zürich

Professor of bioinformatics @ IMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *arraymap* online resource | GA4GH | SPHN

Genomic Background + Disease Parameters

Personalised Medicine
Precision Medicine
...



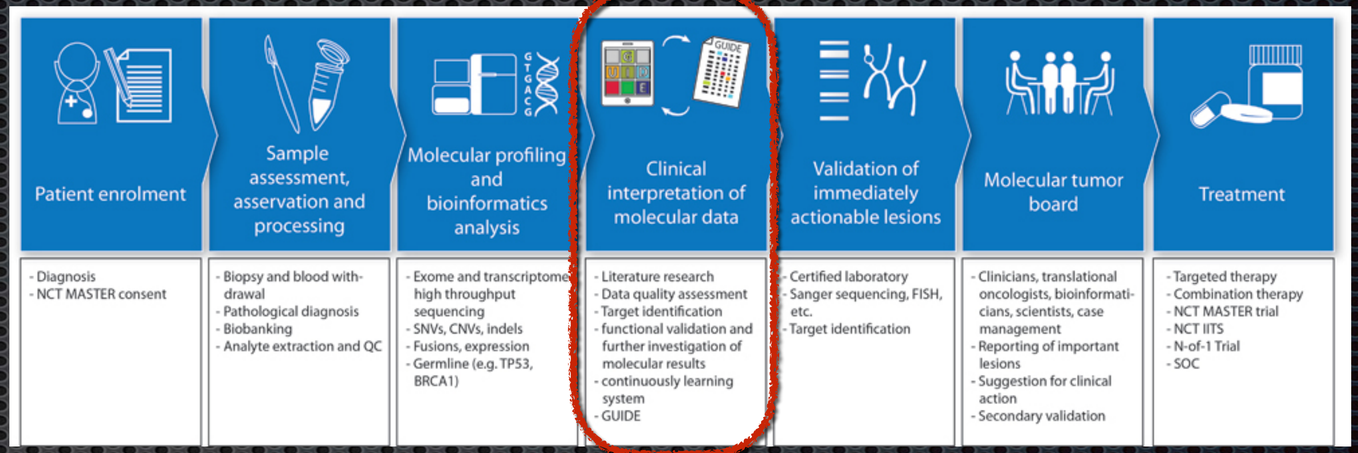


Source: PHG Foundation

While medicine has always been "personal" and "precise" in the given context of available knowledge and technologies, the concept of "**Personalised Medicine**" describes the use of **individual molecular analysis data** and individually **targeted therapies**.

Personalised Medicine in Cancer - A Genome Based Approach

- personalized cancer therapy uses information about the **individual genetic background** and **tumor sequence analysis** for the identification of somatic variants



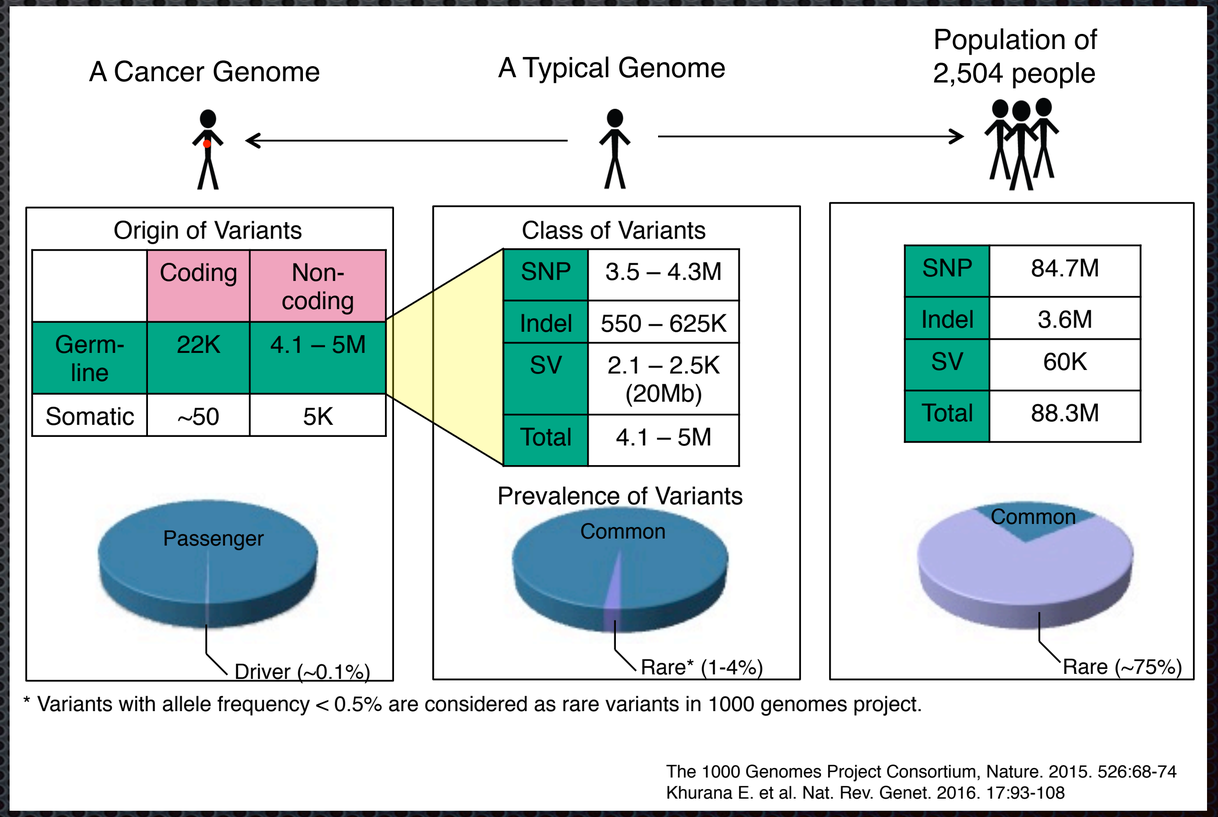
Workflow of a cancer treatment protocol based on "personalized" assessment of actionable genomic lesions (source: NCT Heidelberg).

- currently mostly use of **targeted / panel sequencing** for identification of tens - hundreds of most common "actionable" mutations

- knowledge resources and literature search for interpretation of non-standard variants

Finding Somatic Mutations In Cancer:
Many Needles in a Large Haystack

- a typical humen genome (~3 billion base pairs) has ~5 million variants
- most of them are "rare"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

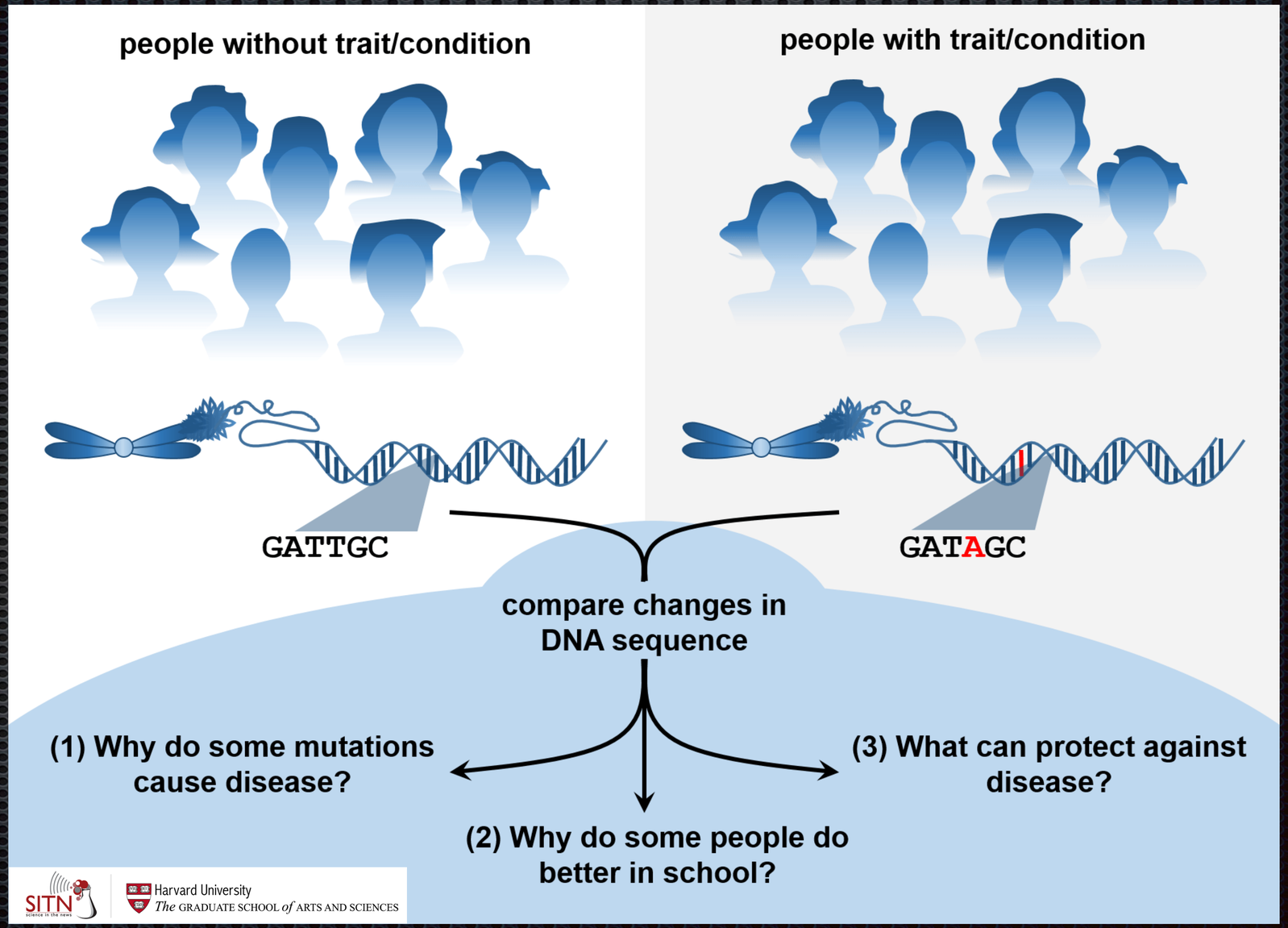


The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

Population Genomics - The Basic Concept

Studying the genomes of many thousands of people at once, known as population sequencing, is an exciting area of research and is producing countless new insights into human biology and medicine. Typically, a population sequencing experiment will involve sequencing the exomes or genomes of large groups of individuals with and without a trait, such as a disease like cancer, then comparing the differences in the genes of the two groups. Once researchers identify DNA changes associated with the trait of interest, they can then use that information to answer many important questions that help guide drug development, clinical practice, and therapeutic selection.



Ryan L. Collins & Brad Wierbowski, Harvard University (2017)

Genomes Everywhere

Large Genome Data Generation, Analysis & Sharing Initiatives

Global Alliance for Genomics & Health

Organization / Initiative: Name	Organization / Initiative: Category	Cohort
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)
23andMe	Organization	>1 million customers (>80% consented to research)
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls
DECIPHER	Repository	19,014 patients (international)
deCode Genetics	Organization	500,000 participants (international)
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts
International Cancer Genome Consortium (ICGC)	Consortium	PCAWG currently data from >8'000 genomes
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects
Resilience Project	Research Project	589,306 individuals
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)
TBResist	Consortium	>2,600 samples
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)
Vanderbilt's BioVU	Repository	>215,000 samples

>1'000'000

100'000

20'000+

17'000+

8'000+

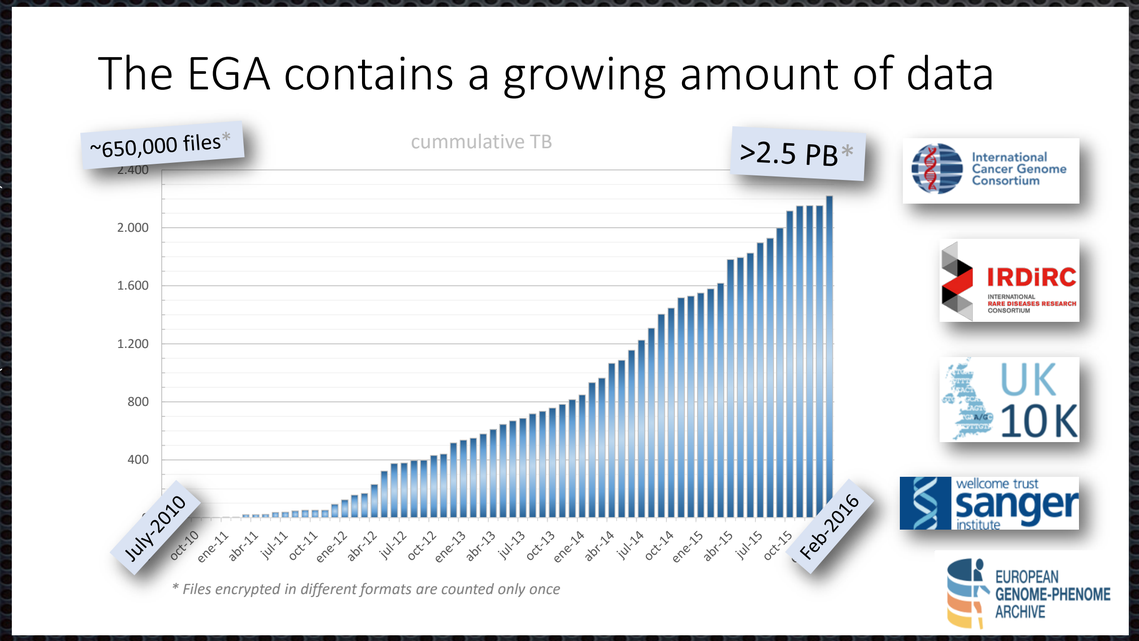
>2-500'000

>1'000'000

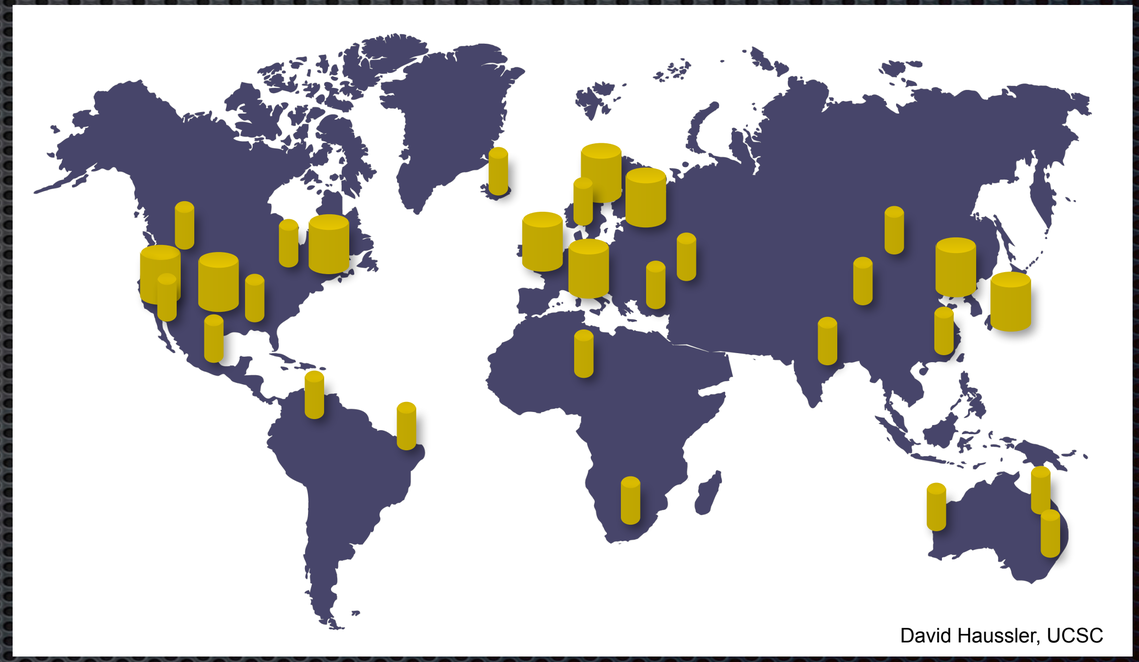
500'000

Genome Datasets: Rapid Growth, Limited Access

population based and cancer research studies produce a rapidly increasing amount of genome sequence data



Source: Jordi Rambla (CRG Barcelona) 2016



genome data is stored in an increasing number of institutional and core repositories, with **incompatible data** structures and **access** policies

Global Alliance
for Genomics & Health

GA4GH to solve genome data access…

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Search...

ABOUT US | HOW WE WORK | GA4GH TOOLKIT | NEWS & EVENTS | BECOME A MEMBER | CONTACT US

Enabling genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework

Genomic Data Toolkit →

Regulatory & Ethics Toolkit →

Data Security Toolkit →

VIEW OUR LEADERSHIP

MORE ABOUT US

BECOME A MEMBER

GA4GH HISTORY & MILESTONES

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary, Orlando
- May 2018 - Vancouver
- October 2018 - 6th plenary, Basel

GA4GH API promotes sharing

GENOMICS

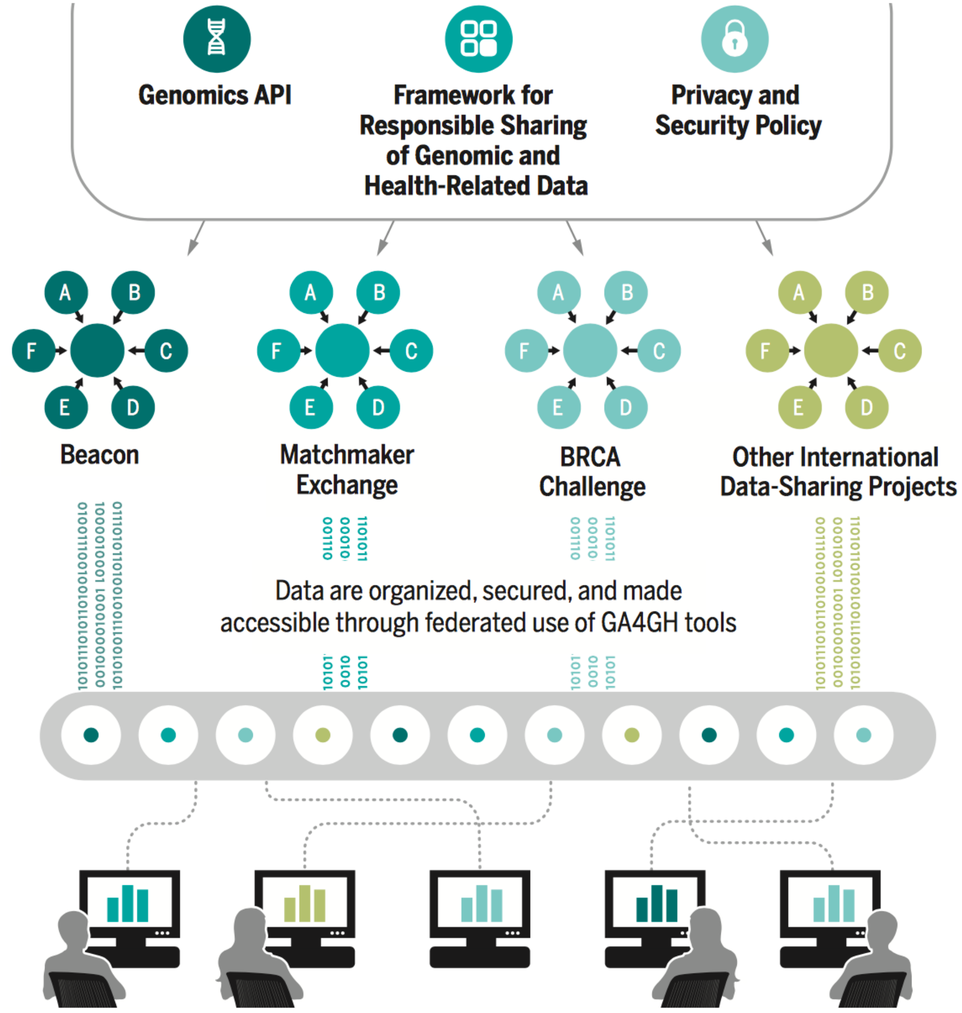
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

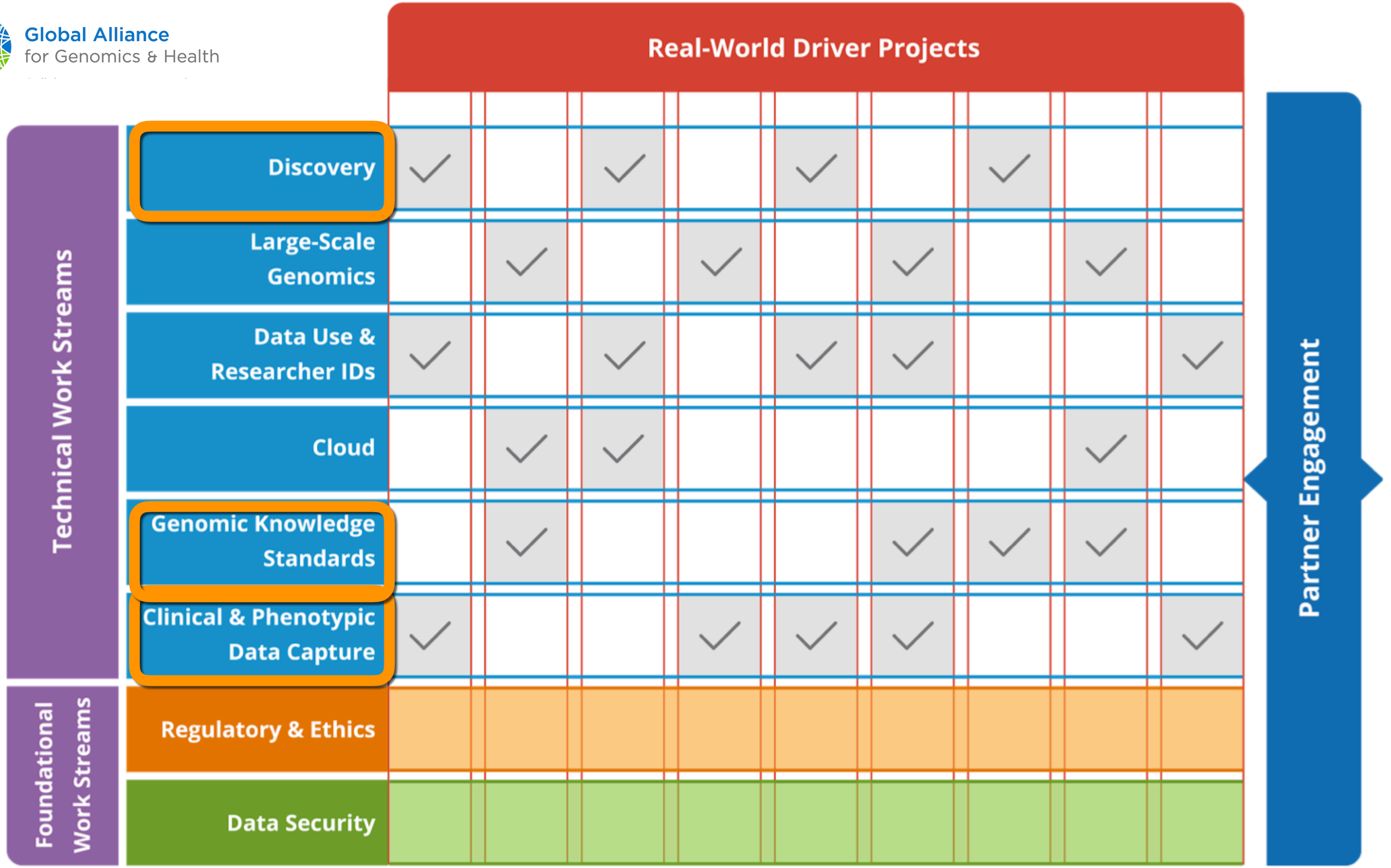
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



Global Alliance for Genomics & Health

Global Alliance
for Genomics & Health



	Role in ELIXIR	Clinical & Phenotypic Data Capture			Cloud						Discovery					Data Use and Researcher Identites		Large Scale Genomics					Genomic Knowledge Standards			Regulatory & Ethics			Data Security	
D - Driver A - Assists C - Consulted I - Informed X - Interested (not yet involved)	P - Platform C - Community IS - Implementation Studies NS - Node Service	Information Models for Clinical/Genomic Data Exchange	Phenotype & Disease Ontology Recommendations	Phenotype Standards Implementation	DOS	TES	WES	TRS	TESTBED	CWL??	Search	Beacon	Service registry	Variant Submission	Internet of Genomics	Data Use Ontology	Researcher Identity and Bona Fide Status	htsget Streaming API	Reference Sequence Retrieval API	RNASeq Expression Matrix	Genetic Variation File Formats	Read File Formats (SAM/BAM/CRAM)	International Code of Conduct for Data Sharing	Variant Annotation: Data Model	Variant Representation: Data Model/Specification	International Participant Values Survey	Return of Results Policy	Cloud Access Policy	AAI	Breach Response Protocol
Belgium																														
Frederik Coppens								I																						
Czech Republic																														
Denmark																														
Jon Ison	P (Tools)							I																						
EMBL																														
Steven Newhouse	P (Compute)				I	I	I	I	I	I	I					I	I	I	I	I	I	I	I				I	I	C	
Susheel Varma	P (Compute)				C	A	A	C	C	I	I		C			C	C	I	I	I	I	I	I				C	C	C	
Thomas Keane	C (Human Data)			D							D	D				D	D	D	D	D	D	D			D				D	
Dylan Spalding	C (Human Data), IS (Beacon)	I	I	D												D	D												D	A
Rob Finn	C (Marine)					I	I	I		I																				
Yasset Perez-Riverol	IS (CWL)					I	I	A		C																				
Michael Crusoe						I	C	I	A	D																				
Guy Cochrane					I					I																				
Sarala Wimalaratne	NS (EMBL-EBI), C(Interop)				C																									
ELIXIR Hub																														
Gary Saunders	C (HDC) IS (Beacon)										C	D	C	I	I															
David Lloyd	C (HDC)										C	D	C	I	I															
Jonathan Tedds	P (Compute)				I	I	I	I	I	I																				
Estonia																														
Jaak Vilo	HoN	I	I	I															I	I	I	I				I	I	I	I	
Hedi Peterson		I	I	I							I	I	I	I	I	I				C	x		I				X		C	
Finland																														
Tommi Nyrönen					I	I	I	I	I	I																				
Shubham Kapoor					C	A	A	C	A	I																				
Ilkka Lappalainen	IS (Beacon)										C	D	C																	
Juha Törnroos	IS (Beacon & CWL)					I	I	I			C	D	C																	
France																														
Francois Moreews								I																						
Olivier Collin								I																						
Genoscope??																														
Roscoff																														
Herve Menager								I																						
David Salgado (would like to be involve	C (hCNV)	I	I	I							A	A	I	I				I	I	I	A	I	I	I	I					
Germany																														
Björn Grüning						I	I	I																						
Alexsander Goesmann						I	I	I																						
Chris Lawerenz (would like to be involv	P (Compute), C	X (I)				X (I)	X (I)	X (I)					X (I)					X (I)	X (I)				X (I)					X (I)	X (I)	X (I)
Greece																														
Hungary																														
Ireland																														
Israel																														
Italy																														
Giacinti Donvito						I	I	I																						
Luxemborg																														
Venkata Satagopam	Beacon											C																		
Netherlands																														
Morris Swertz	P / C / IS (Beacon)	x			x	x	x	x	x	x	a	a		x				x	x		x	x		x					i	
Marco Roos	P / C / IS (Rare Disease)																													
Jildau Bouwman	C (Nutrition)	x		x												x					x			x					
Chris Evelo	P (Interoperability)																													
Norway																														
Elvind Hovig								I		I																				
Portugal																														
Slovenia																														
Spain																														
Jordi Rambla	C (Human Data)										C	D	C																	
Sabela de la Torre	IS (Beacon)										C	D	C																	
Salvador Capella								I																						
Sergi Beltran	C (Rare Diseases)										A	X			X															
Ivo Gut	C (Rare Diseases)										A	I																		
Sweden																														
Niclas Jarenbourg	Beacon											C																		
Switzerland																														
Michael Baudis	IS (Beacon)	D	C	C							C	D	C												C	D			I	
Torsten Schwede						I	I	I																						
Kevin Sayers						I	I	I																						
United Kingdom																														
Mike Sternberg	P (Compute)					I	I	I																						
Stian Soiland-Reyes	P (Interop)					I	I	I		C																				

Global Alliance
for Genomics & Health

GA4GH Driver Projects

BRCA Challenge
The BRCA Challenge aims to advance understanding of the genetic basis of breast and other cancers using data from around the world.

Beacon Project
Beacon Project is an open web service that tests the willingness of international sites to share genetic data. It is being implemented on the websites of the world's top genomic research organizations.

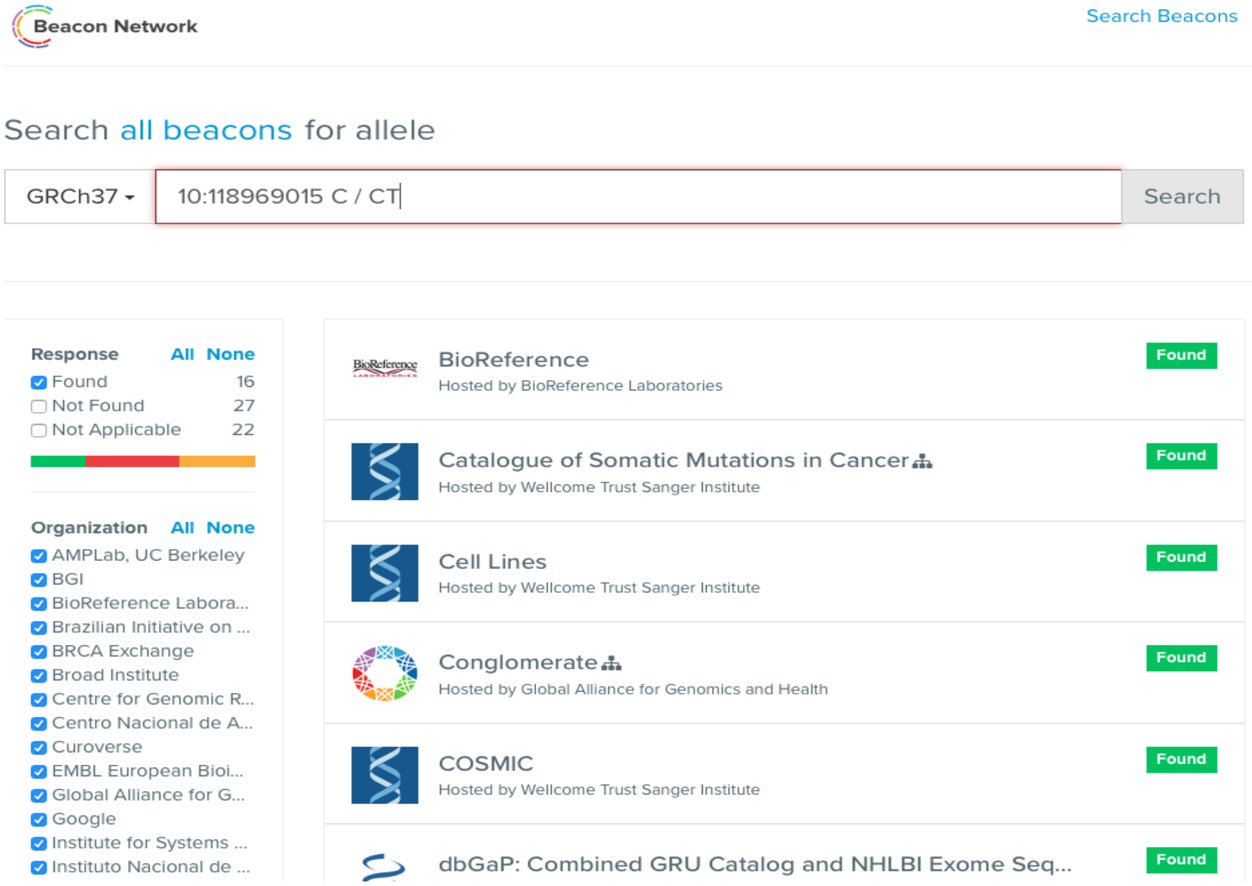
Matchmaker Exchange
Matchmaker Exchange is a federated network of databases whose goal is to find genetic causes of rare diseases by matching similar phenotypic and genotypic profiles.

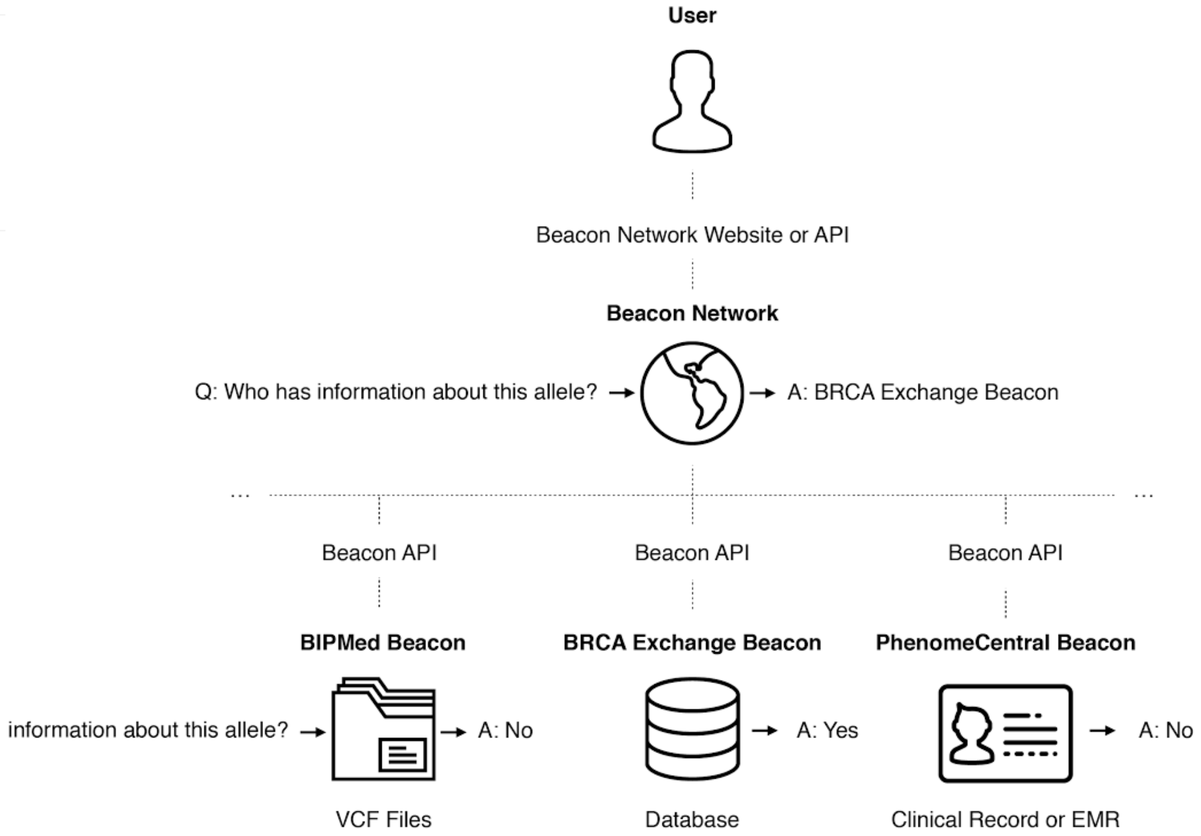
BRCA
CHALLENGE

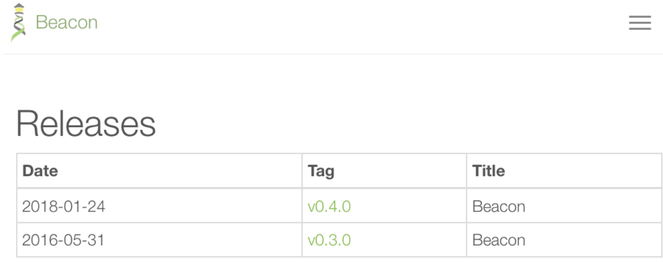
Matchmaker
Exchange

Beacon Project

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







Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

200+ Datasets

100K+ Individuals

baudisgroup @ UZH & SIB

Computational Cancer Genomics | Data Sharing | Populations Background | Tools

University of Zurich[UZH]

Global Alliance for Genomics & Health

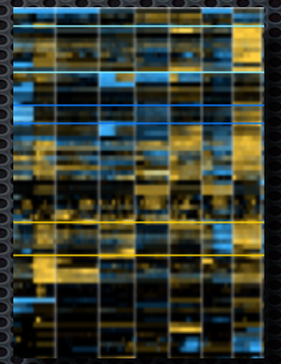
baudisgroup @ UZH & SIB

beyond cancer genomes

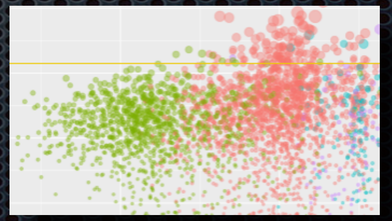
- Cancer genome data resources
- Software tools for data analysis & visualisation
- Parsing the cancer genome landscape: Patterns & targets
- GA4GH: **Standards** & **Beacons**
- Quantifying cancer genomic research
- The Swiss Personalized Health Network initiative
- Collaborations!

progenetix

arrayMap



Global Alliance for Genomics & Health



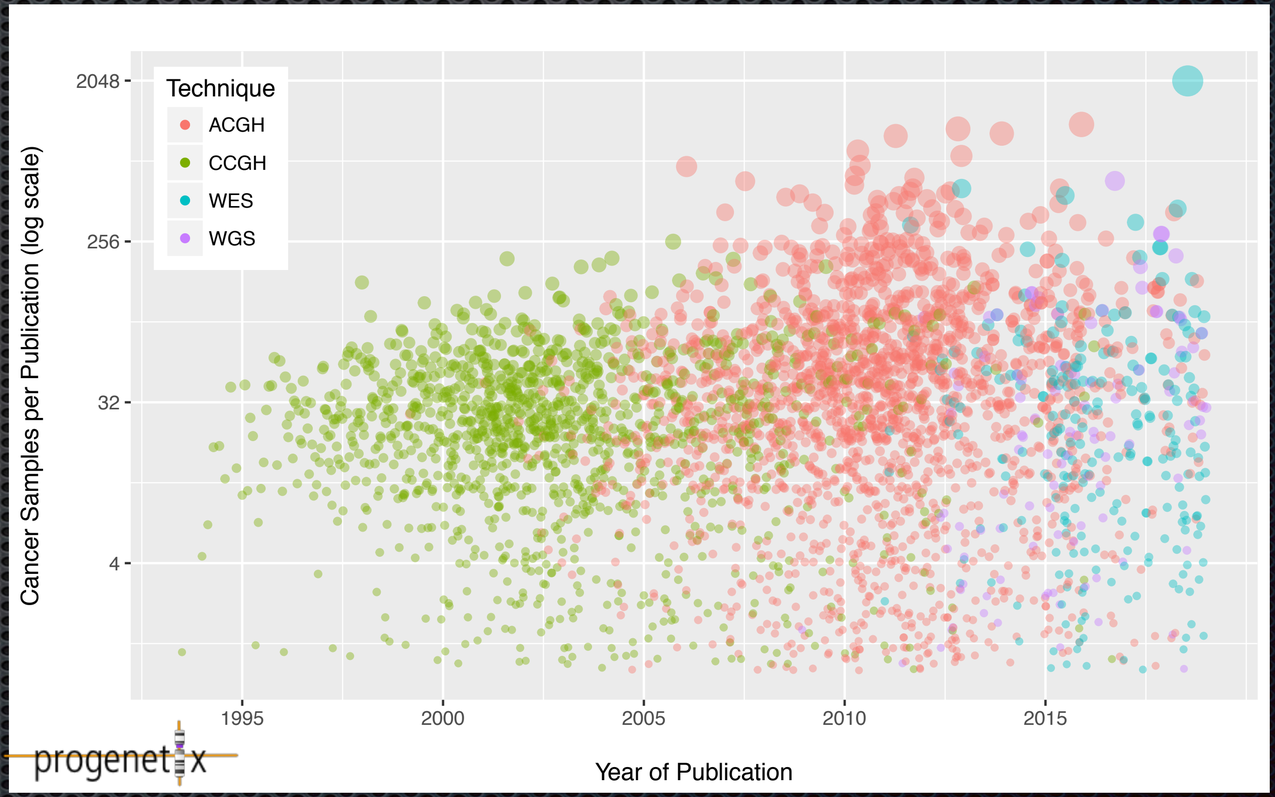
SAMW ASSM

akademien der wissenschaften schweiz

SIB Swiss Institute of Bioinformatics

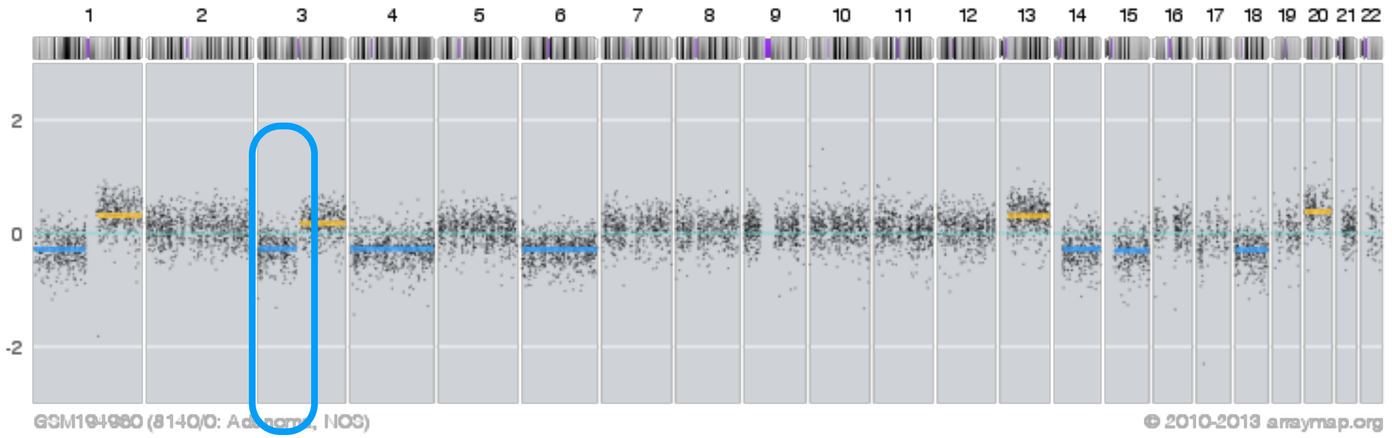
Molecular Cytogenetic & Sequencing Studies for **Whole Genome Profiling** in Cancer

- genome screening to identify mutations in cancer samples
- for diagnostic purposes and therapeutic target identification
  - karyotyping (~1968)
  - **C**omparative **G**enomic **H**ybridization (1992)
  - genome **microarrays** (aCGH, SNP arrays ...; 1997)
  - **W**hole **E**xome **S**equencing (2010)
  - **W**hole **G**enome **S**equencing (2011)

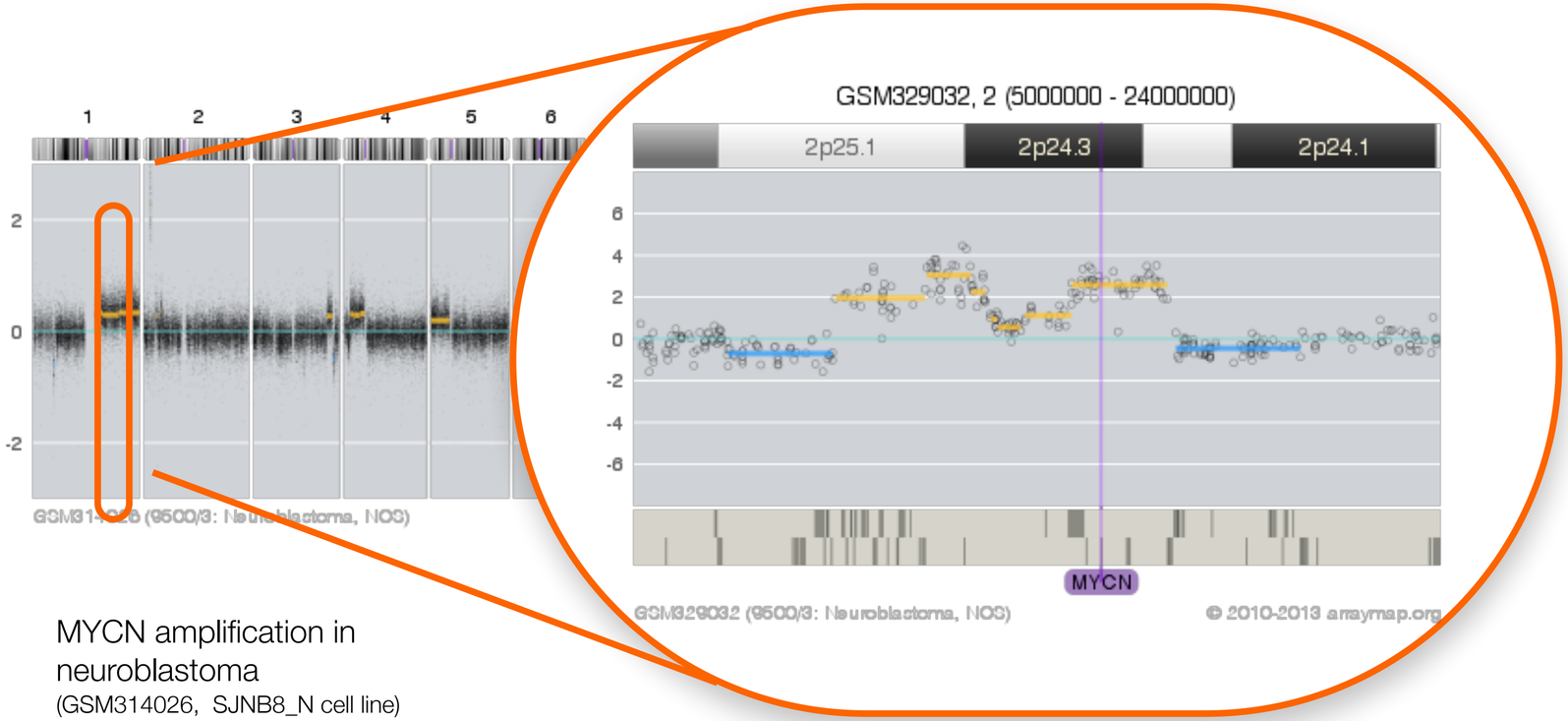


Overview of publications reporting whole-genome screening analysis of cancer samples, by molecular-cytogenic or genome sequencing methods. The data represents 3357 articles assessed for the progenetix.org cancer genome data resource (M. Baudis, 2001-2019

Copy number aberrations



Gain of chromosome arm 3q in colorectal carcinoma

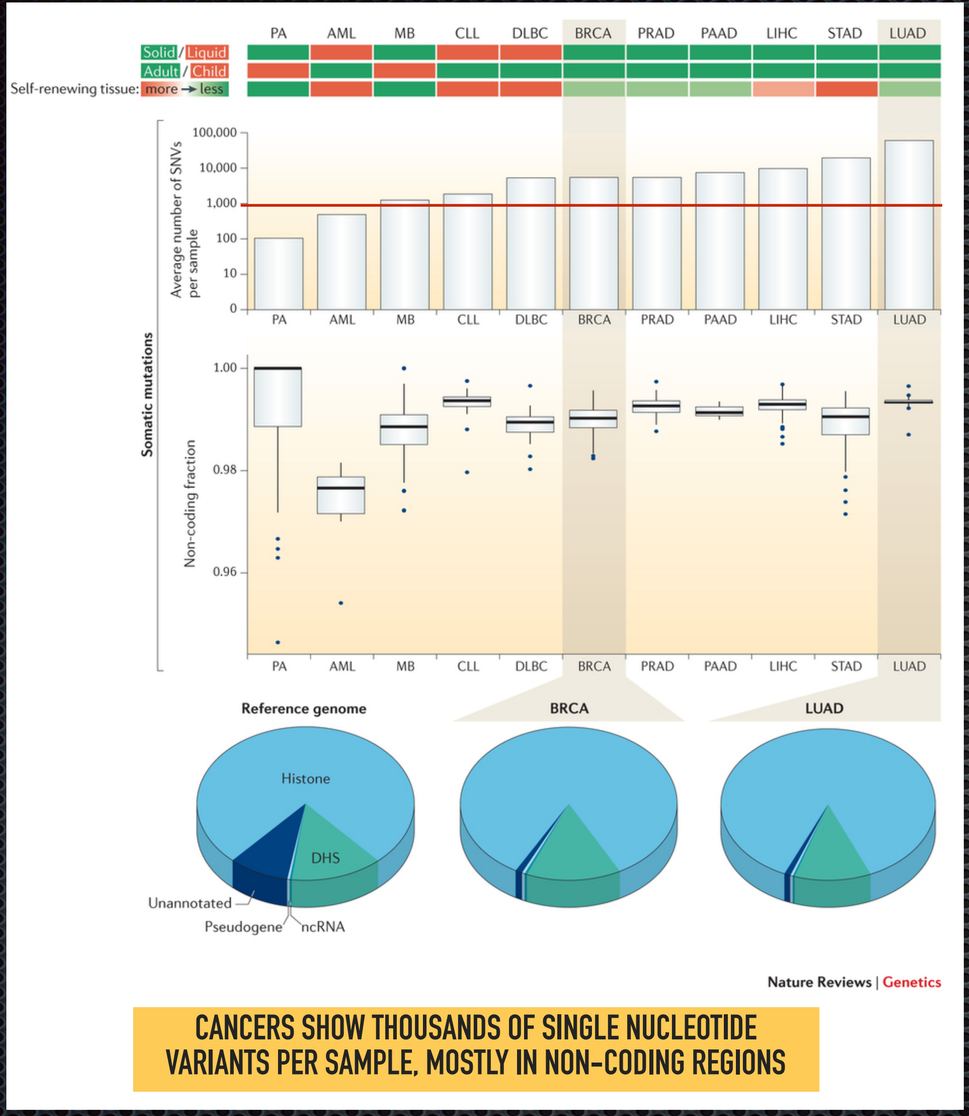


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

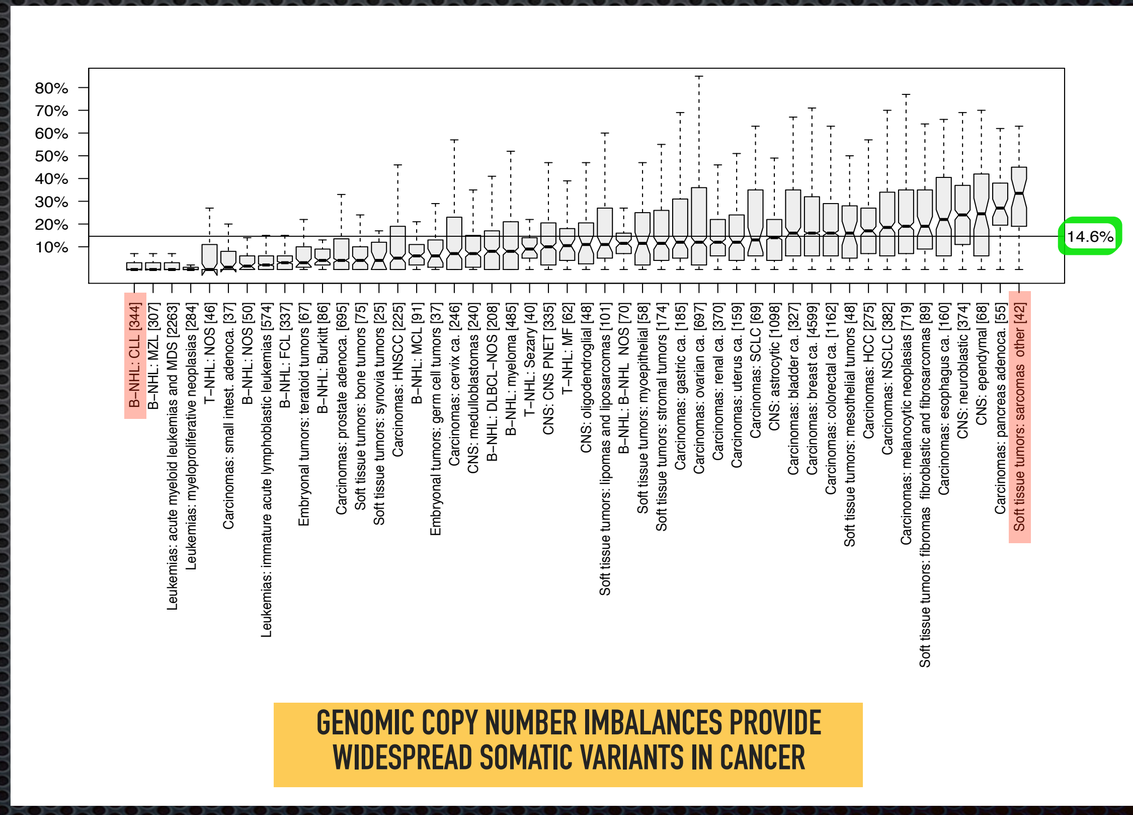
low level/**high level** copy number alterations (CNAs)

arrayMap

Quantifying Somatic Mutations In Cancer



CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS



GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

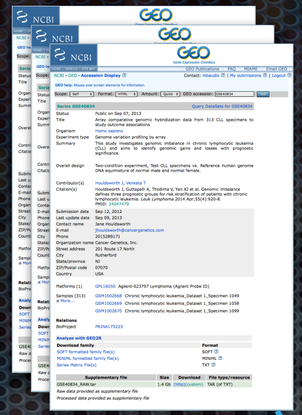
On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on >30'000 cancer genomes from arraymap.org

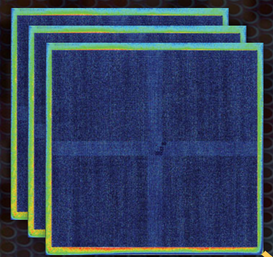
Reference Resources for Cancer Genome Profiling

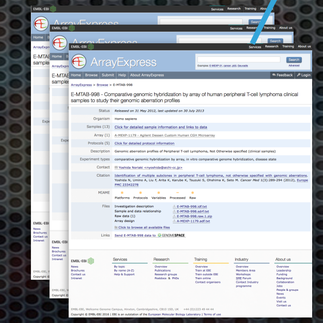
- continuously updated reference resources for cancer genome profiling data and related information
- basis for own research activities, collaborative projects and external use
- structured information serves for implementing GA4GH concepts

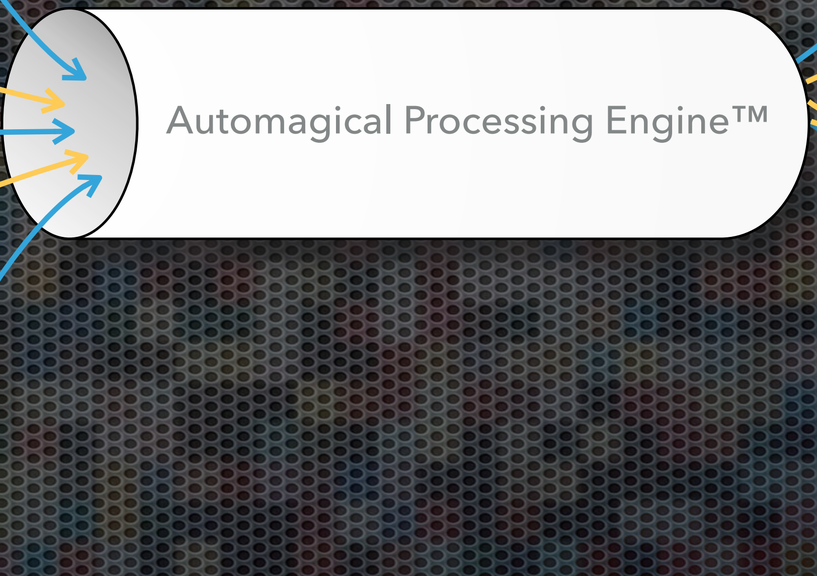
	progenetix	arrayMap
techniques	cCGH, aCGH, WES, WGS	aCGH (+?)
scope	sample (e.g. combination of several experiments)	experiment
content	**>31000** samples	**>60000** arrays
raw data presentation	no (link to sources if available)	yes (raw, log2, segmentation if available)
per sample re-analysis	no; supervised result (mostly as provided through publication)	yes (re-segmentation, thresholding, size filters ...)
final data	annotated/interpreted CN status for GP and cytogenetic regions	unsupervised CN status for GP and cytogenetic regions
main purposes	• Distribution of CNA target regions in most tumor types (>350 ICD-O) • Cancer classification	• Gene specific hits • Genome feature correlation (fragile sites ...)

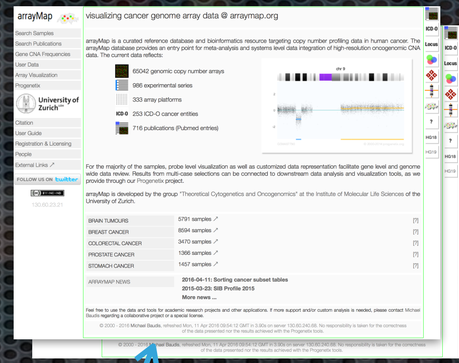
DATA PIPELINE

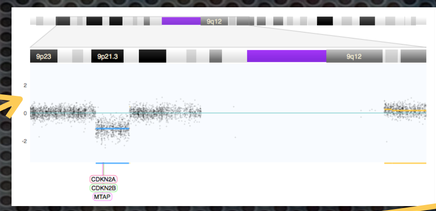


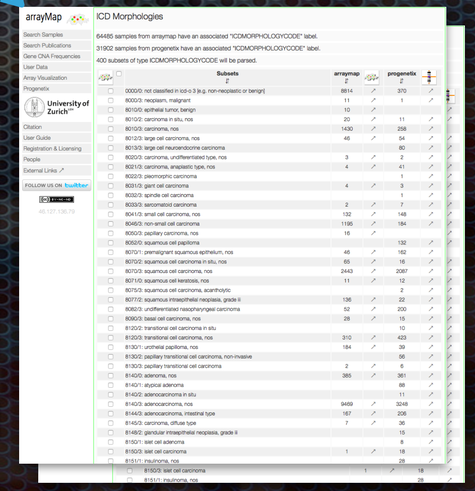


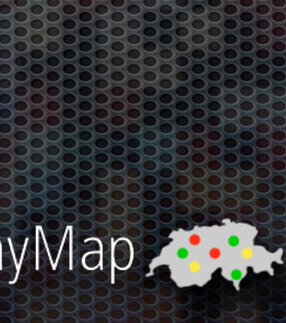


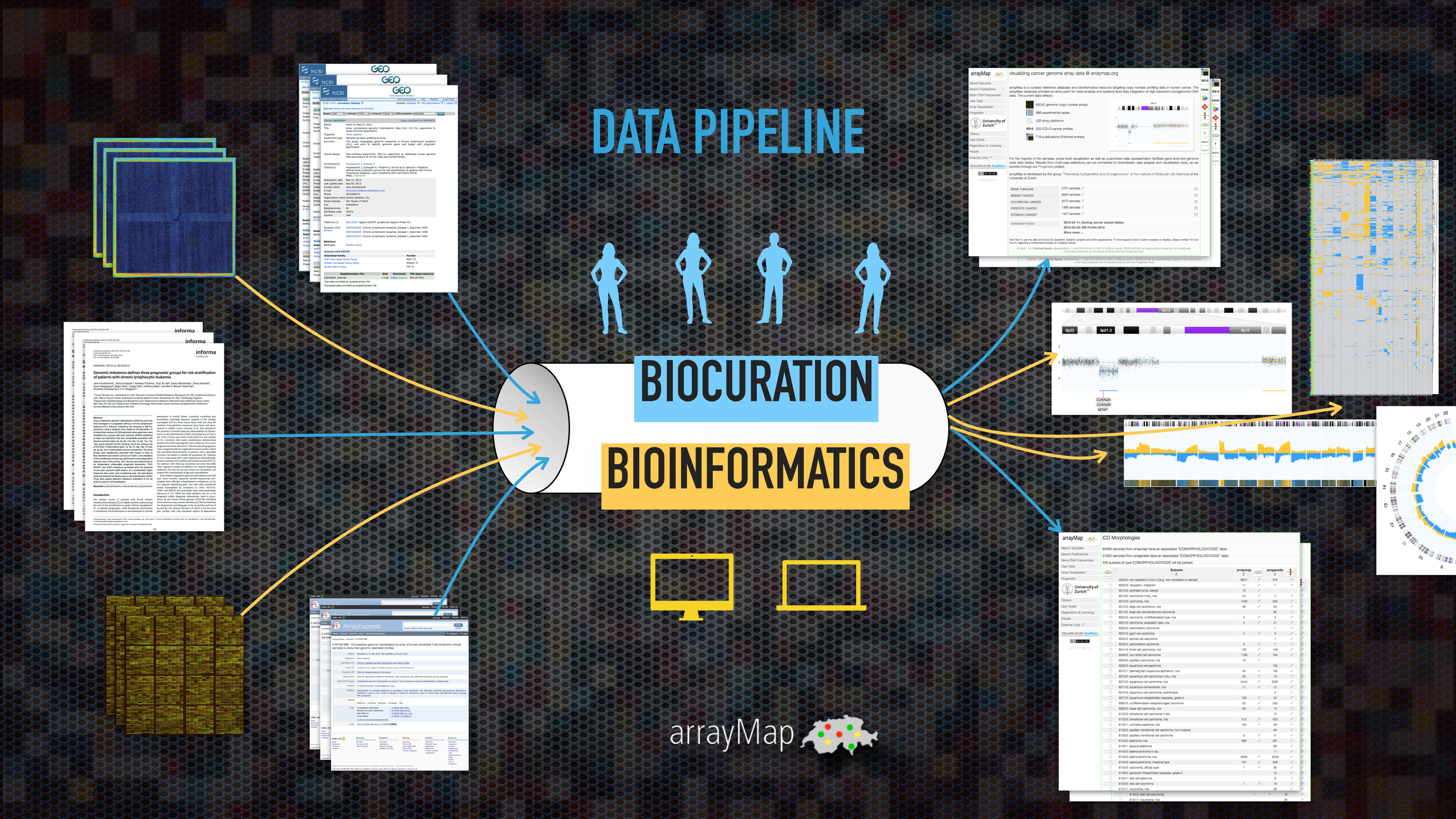











DATA PIPELINE
BIOCURATION
BIOINFORMATICS
arrayMap
visualizing cancer genome array data @ arraymap.org
ICD Morphologies
Genomic imbalance defines three prognostic groups for risk stratification of patients with chronic lymphocytic leukemia

ARRAYMAP DATA "PIPELINE"

GEO
ArrayExpress

amDownload

Metadata
(GEO soft files…)

amParseMeta

metadata.ods
(multi-sample)

amUpdateMeta

db.metadata

Array Probe Data
(log2
tables, .CEL)

Metadata Review
Selection
Entity Mapping
(ICD-O, Biosample,
Individual, clinical...)

db.experiments

Publications
Direct
Submissions

Array
Processing
Pipeline

Manual interpretation
of associated
information (article,
series descriprtion...)

Array
Processing
Pipeline

arrayMap

Batch Tool for Genome Liftover

Challenge:

1. Keep the integrity of copy number segments after Liftover.
2. 10% data lost from straight Liftover.
3. 1TB segment and probe data in over 2,000 nested directories

Solution:

1. Algorithm to lift segments.
2. Algorithm for fuzzy remapping.
3. Parallel processing and failure recovery mechanism

F1000Research 2018, 7:319 Last updated: 08 JUN 2018

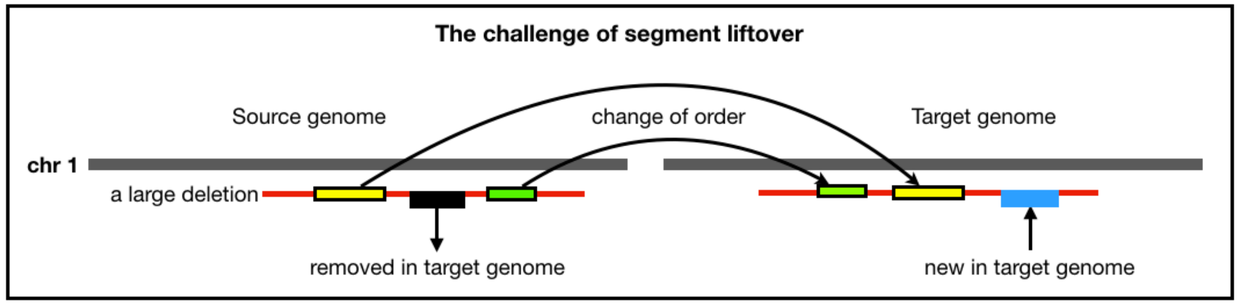
Check for updates

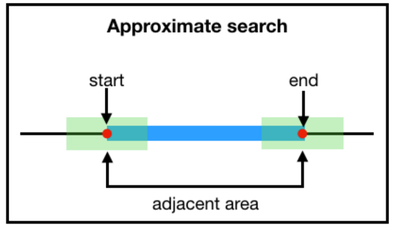
SOFTWARE TOOL ARTICLE

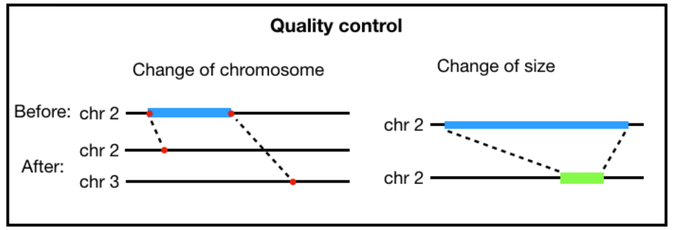
REVISED **segment_liftover : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]**

Bo Gao [1,2], Qingyao Huang[1,2], Michael Baudis [1,2]

[1]Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland
[2]Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland



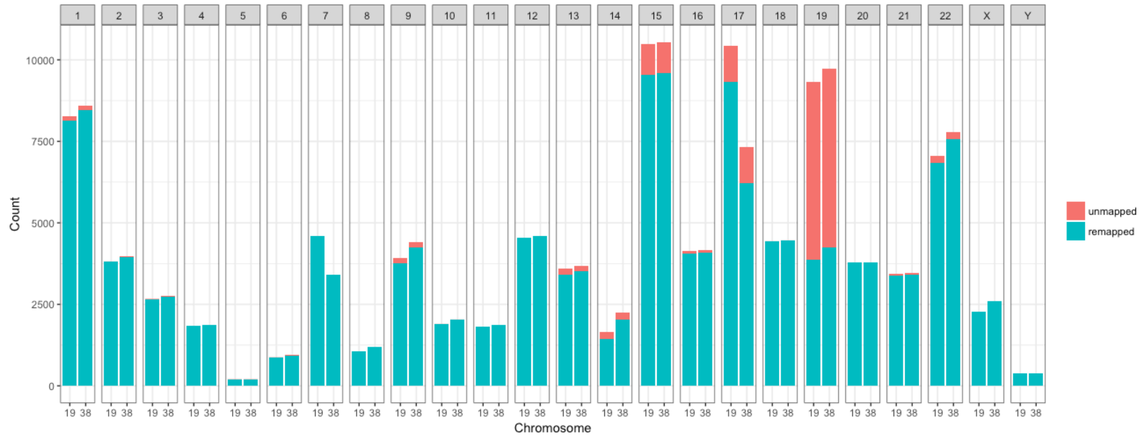


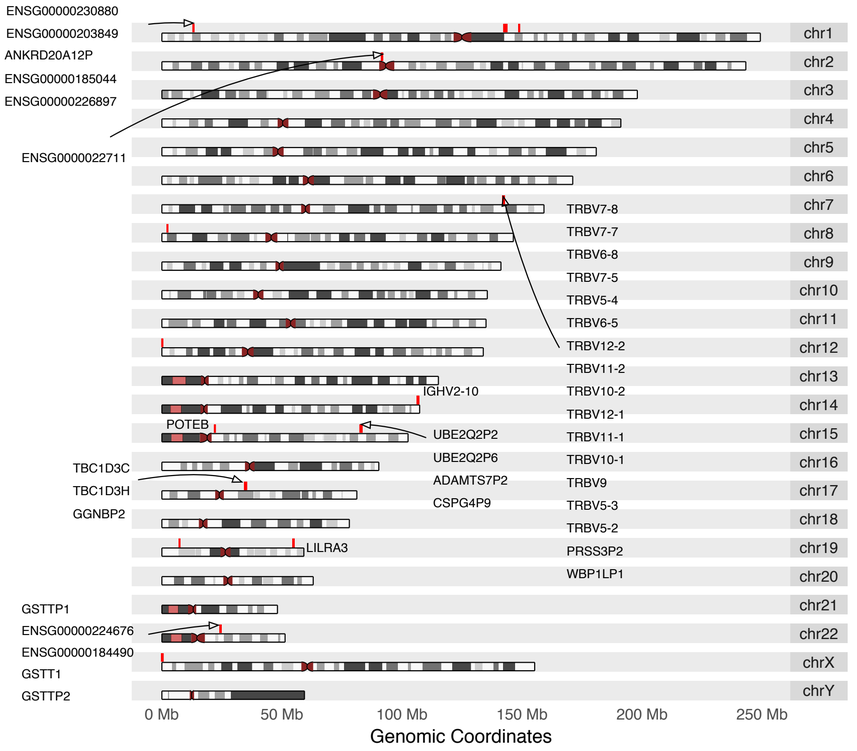


Results of Liftover

- Converted 44,632 probe files and 44,471 segment files from hg19 to hg38.
- The probe data were generated from nine Affymetrix genotyping array platforms, which currently only support annotations for hg19.
- It took 42 hours to convert 5.5 billion probes and 40 minutes to convert 4.8 million segments. (12-core, 128GB RAM machine with 8 parallel processes)
- Reduced information lost from 10% to less than 0.1%

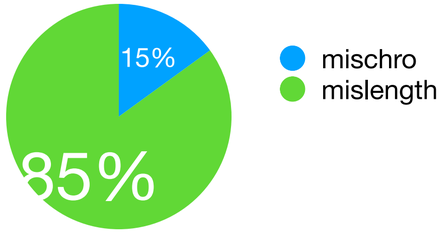
Distribution of remapped and unmapped





Distribution of unmapped probes

Reason of unmapped segments



arrayMap

Resource for copy number variation data in cancer

arrayMap

- Search Samples
- Search Publications
- Gene CNA Frequencies
- User Data
- Array Visualization
- Progenetix

University of Zurich[UZH]

- Citation
- User Guide
- Registration & Licensing
- People
- External Links ↗

FOLLOW US ON twitter

(cc) BY-SA

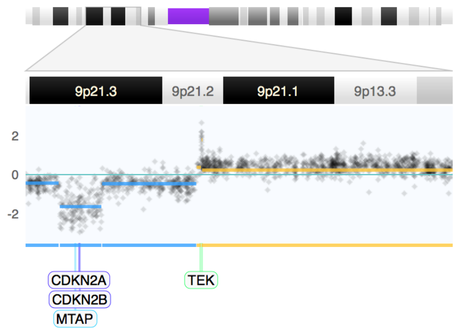
130.60.23.21

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 63060 genomic copy number arrays
- 763 experimental series
- 145 array platforms
- ICD-O 141 ICD-O cancer entities
- 554 publications (Pubmed entries)



Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma (**GSM491153**), indcating, among others, a homozygous deletion involving CDKN2A/B.

ICD-O
Locus
?
HG18
HG19

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

BRAIN TUMOURS	5653 samples ↗	[?]
BREAST CANCER	8329 samples ↗	[?]
COLORECTAL CANCER	3238 samples ↗	[?]
PROSTATE CANCER	991 samples ↗	[?]
STOMACH CANCER	1062 samples ↗	[?]

ARRAYMAP NEWS

2016-08-03: SVG graphics

2016-05-17: Transitioning to Europe PMC

More news ...

Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project or a special license.

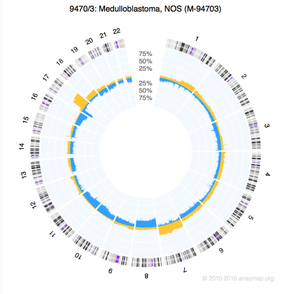
ICD Morphologies

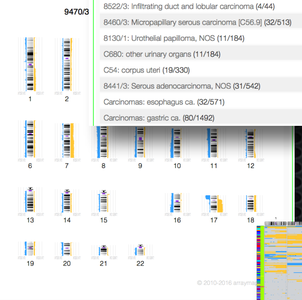
2021 samples from arraymap have an associated "ICDMORPHOLOGYCODE" label.

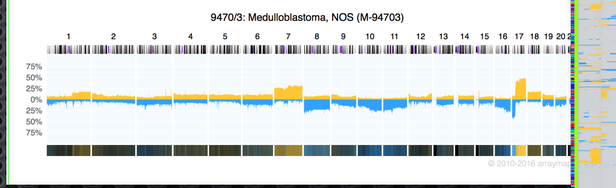
9470/3: Medulloblastoma, NOS (M-94703)

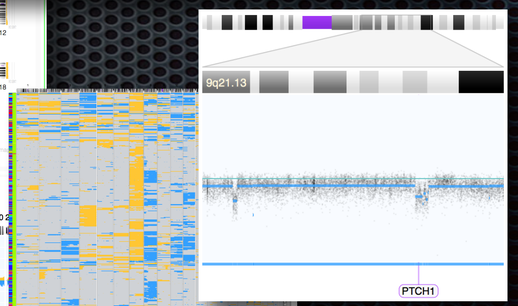
Synonyms

- Medulloblastoma, NOS
- Melanotic medulloblastoma









FIND CNAS BY GENE OR REGION: TP53 | [ERBB2] 17:35097862-35138441:1 [?]

REGION SIZE | MAX COVERAGE (KB): 0 kb - 5000 - 250000 kb [?]

CLINICAL DATA: no followup required

CITY: km [?]

Query Database

1949 of 65042 cases matched the selection criteria.

SUBSET	PERCENT IN SUBSET
8507/3: Invasive micropapillary carcinoma (13/39)	33.3
C692: retina (14/82)	17.1
8260/3: Papillary adenocarcinoma, NOS (11/65)	16.9
8500/3: Invasive carcinoma of no special type (1201/8188)	14.7
8560/3: Adenosquamous carcinoma (3/21)	14.3
Carcinomas: breast ca. (1254/8837)	14.2
C50: breast (1254/8929)	14.0
8500/2: Ductal carcinoma in situ, NOS (25/225)	11.1
C32: larynx (3/29)	10.3
8010/2: Carcinoma in situ, NOS (2/20)	10.0
C187: sigmoid incl. rectosigmoid junction (13/140)	9.3
8480/3: Mucinous adenocarcinoma (12/132)	9.1
8522/3: Infiltrating duct and lobular carcinoma (4/44)	9.1
8460/3: Micropapillary serous carcinoma [C56.9] (32/513)	6.2
8130/1: Urothelial papilloma, NOS (11/184)	6.0
C680: other urinary organs (11/184)	6.0
C54: corpus uteri (19/330)	5.8
8441/3: Serous adenocarcinoma, NOS (31/542)	5.7
Carcinomas: esophagus ca. (32/571)	5.6
Carcinomas: gastric ca. (80/1492)	5.4

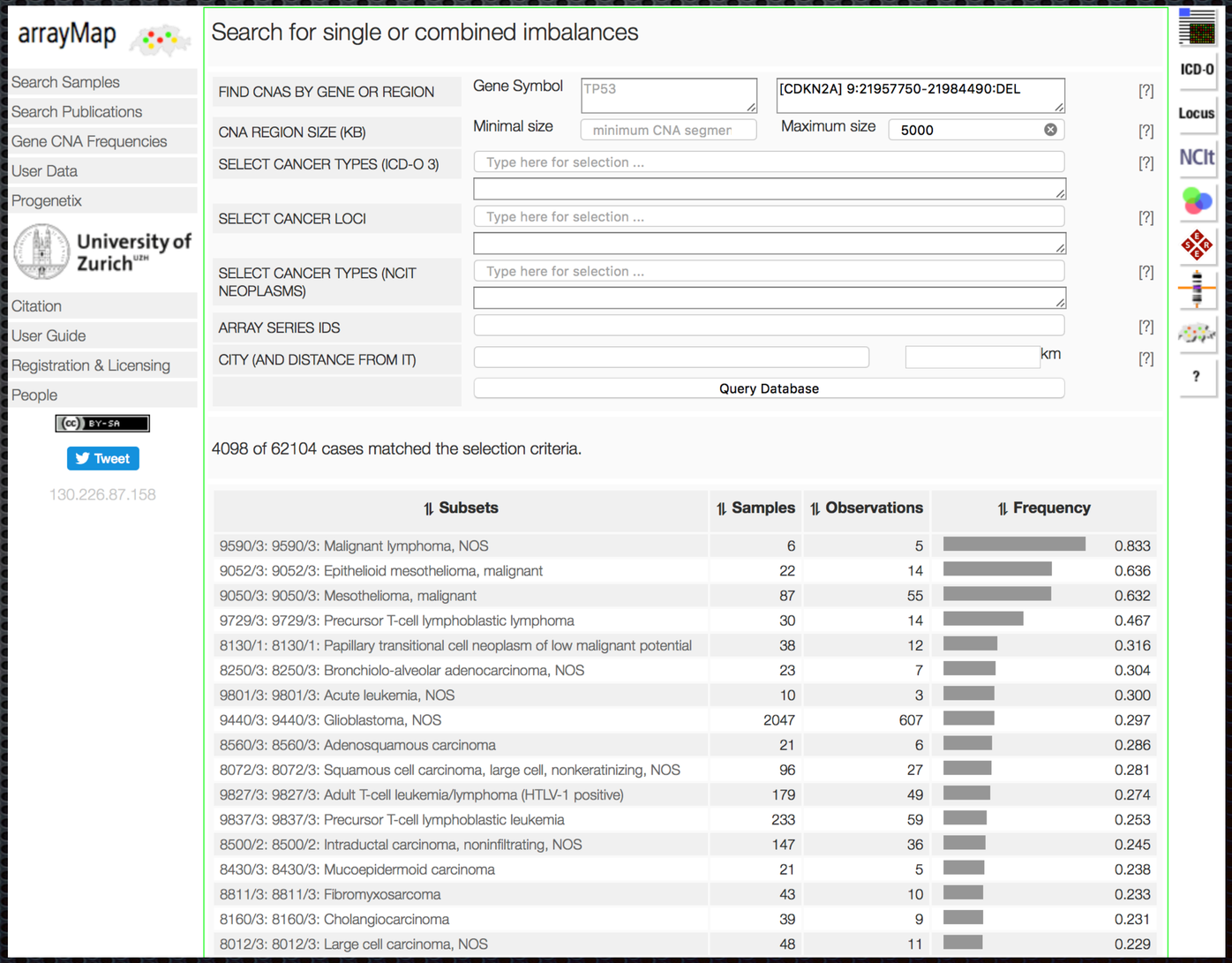
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	GSE40834	24047479	9823/3	C42
GSM1002669	GSE40834	24047479	9823/3	C42
GSM1002670	GSE40834	24047479	9823/3	C42
GSM1002671	GSE40834	24047479	9823/3	C42
GSM1002672	GSE40834	24047479	9823/3	C42
GSM1002673	GSE40834	24047479	9823/3	C42
GSM1002674	GSE40834	24047479	9823/3	C42
GSM1002675	GSE40834	24047479	9823/3	C42
GSM1002676	GSE40834	24047479	9823/3	C42
GSM1002677	GSE40834	24047479	9823/3	C42
GSM1002678	GSE40834	24047479	9823/3	C42
GSM1002679	GSE40834	24047479	9823/3	C42
GSM1002680	GSE40834	24047479	9823/3	C42

arrayMap

arrayMap

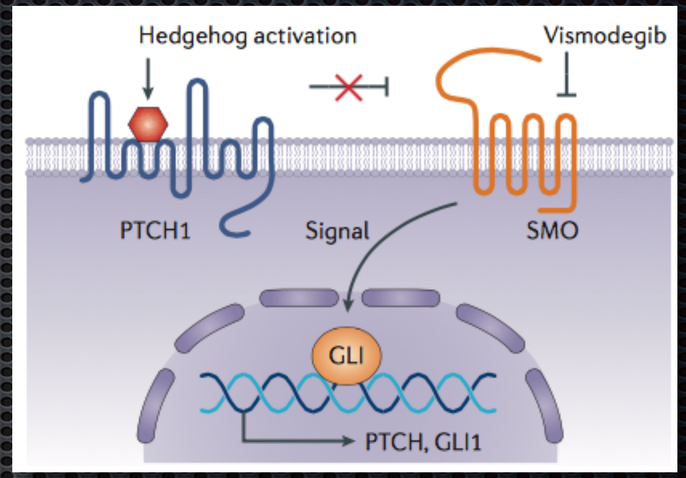
Resource for copy number variation data in cancer

- Typical use: Gene CNA frequencies
- allows for (gene symbol supported) query for DUP/DEL overlapping a region
- size windowing ("focal" changes)
- no thorough statistics implemented yet (e.g. modeling overall CNA background)

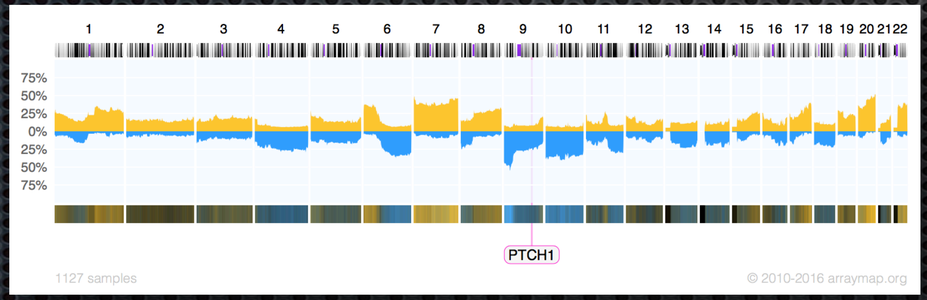


Rare Events & Hidden Therapeutic Options?
Example: PTCH1 deletions in malignant melanomas

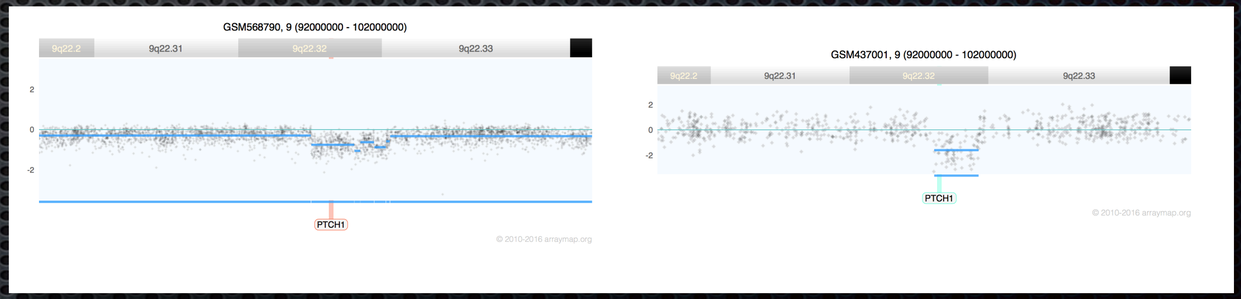
- PTCH1 is a actionable tumor suppressor gene, which has been demonstrated in e.g. basaliomas and medulloblastomas
- analysis of 1127 samples from 26 different publications could identify **focal** deletions in 4 samples
- a current project addresses the focal involvement of all mapped genes, in >50'000 cancer genome profiles



In its normal function, PTCH1 is a tumor suppressor gene in the sonic hedgehog pathway and inhibits SMO driven transcriptional activation. A loss of PTCH1 function (mutation, deletion) can be mitigated through drugs antagonistic to SMO activation.



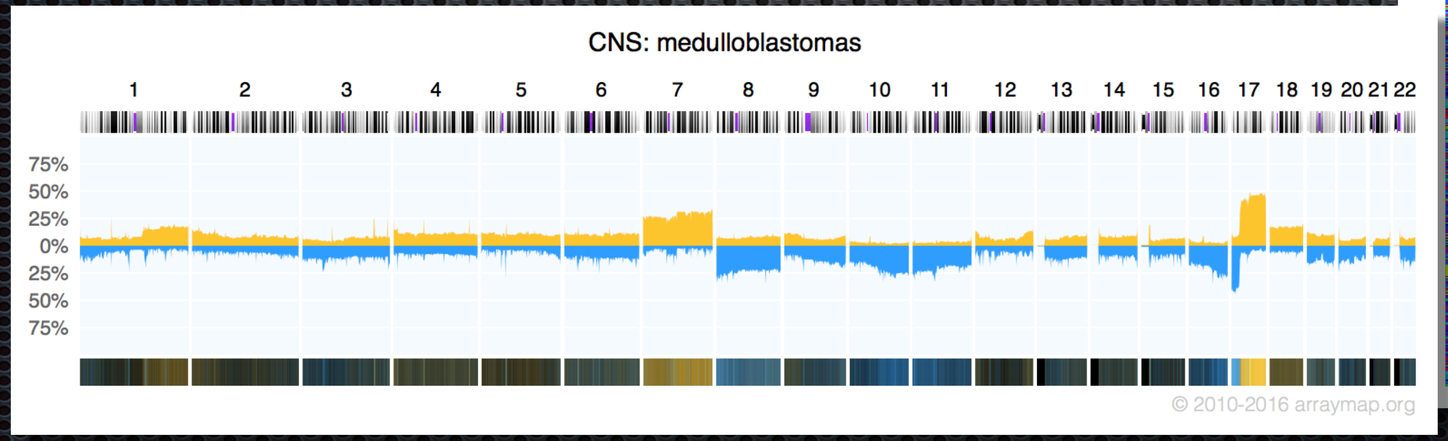
Summary of somatic copy number aberrations from the analysis of 1127 genome profiles of malignant melanomas, collected in our arraymap.org cancer genome resource. While PTCH1 does not represent a deletion hotspot, the genomic locus is part of larger deletions in ~25% of melanoma samples.

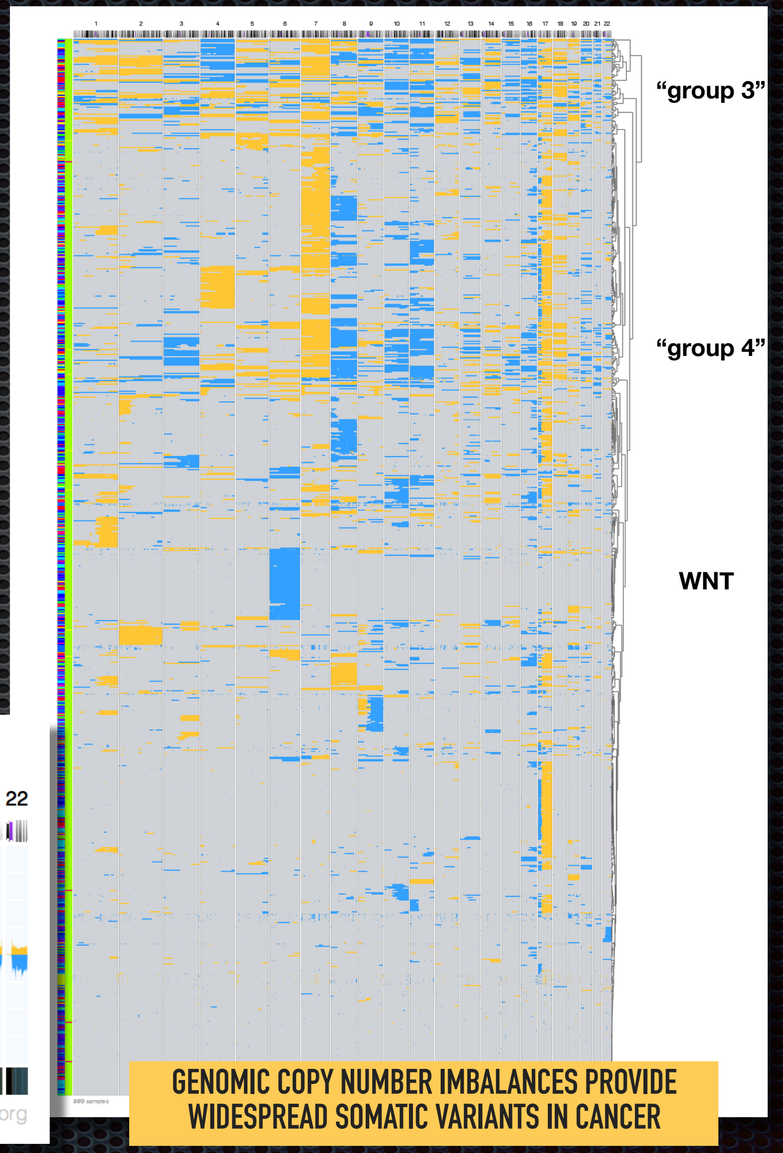


Examples of focal / homozygous PTCH1 deletions detected in the analysis of 1127 genomic array datasets. Focal somatic imbalance events are considered an indicator for oncogenic involvement of the affected target genes.

Somatic Mutations In Cancer: Patterns

- many tumor types express **recurrent mutation patterns**
- **How can** those patterns be used for classification and determination of biological mechanisms?





A genomic copy number histogram for malignant medulloblastomas, the most frquent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional relevance and clinical correlation. From arraymap.org

Somatic Mutations In Cancer: Patterns III
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles

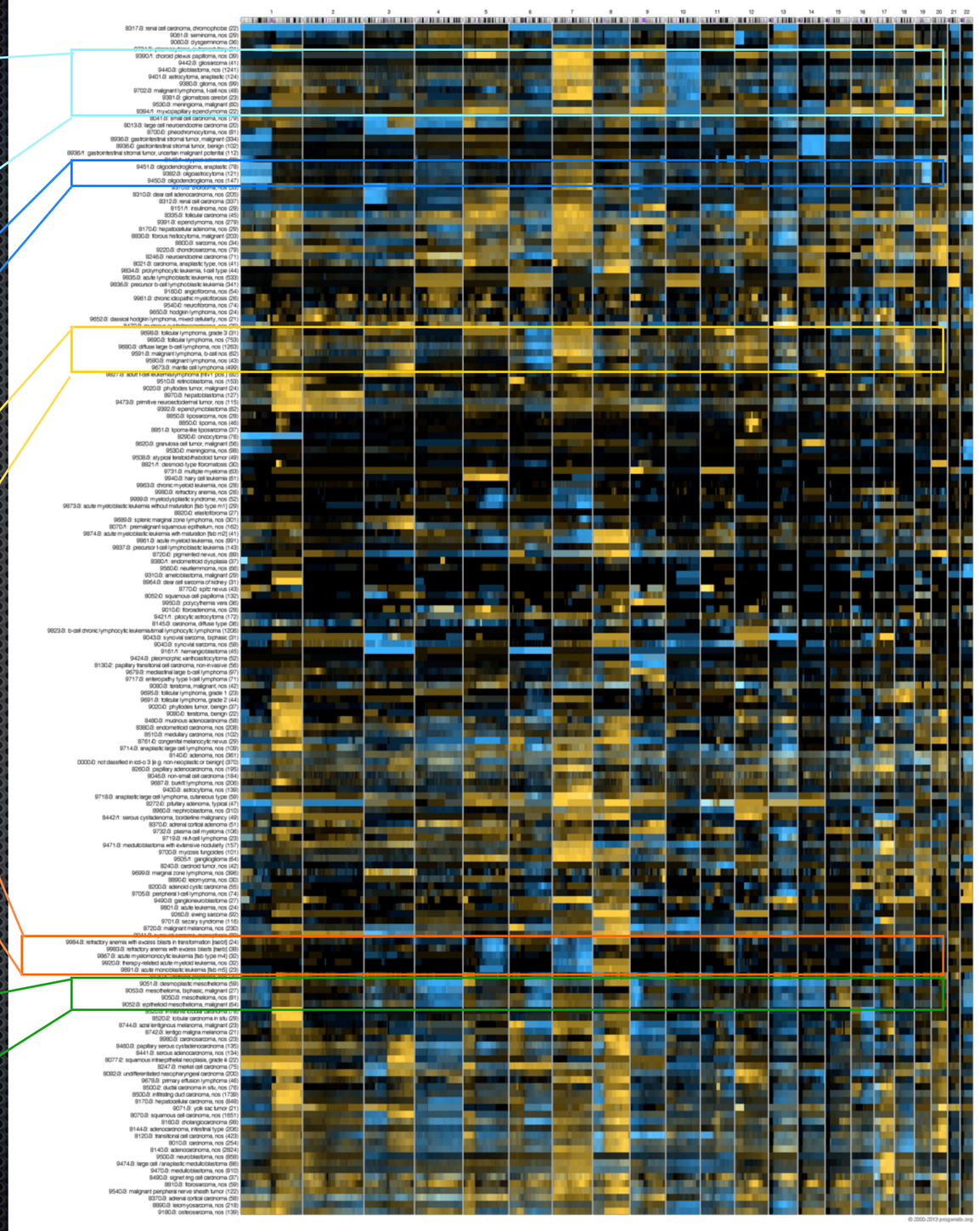
9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)



arrayMap

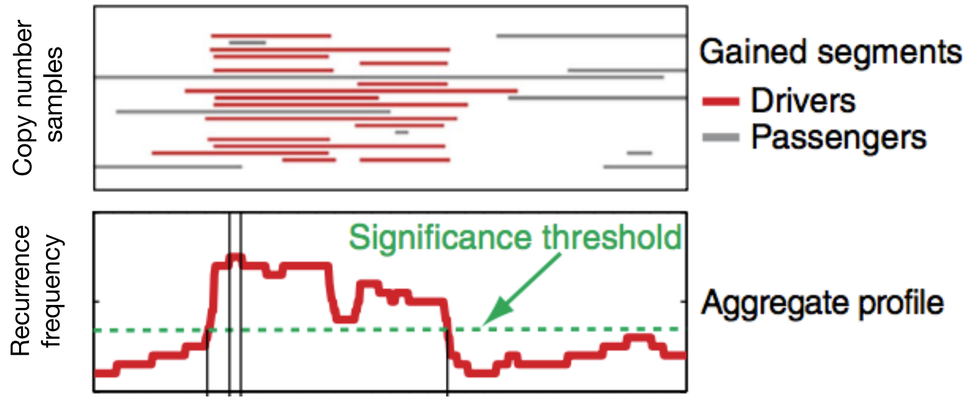
A machine learning approach to identify driver genes in cancer development

Traditional approach

- method: finding focal signal peaks
- data: point mutation or copy number variation

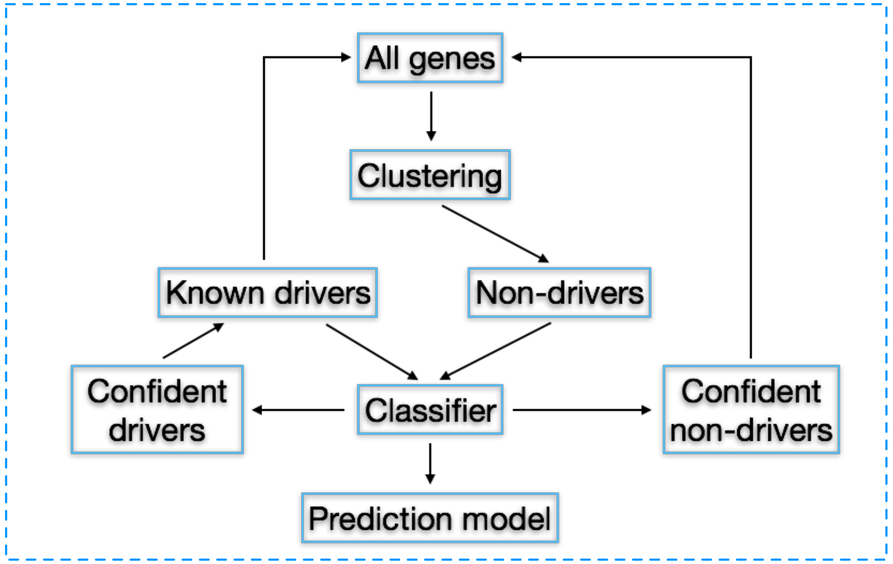
New approach

- method: unsupervised machine learning
- data: point mutation, copy number variation and methylation



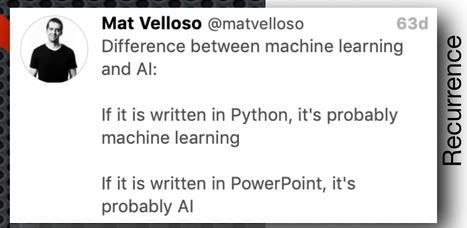
The classical approcah of driver identificaiton

Cited from: Van Dyk, Ewald, et al. *Nature communications* 7 (2016).



The recursive learning workflow

An **Artificial Intelligence** approach to identify driver genes in cancer development

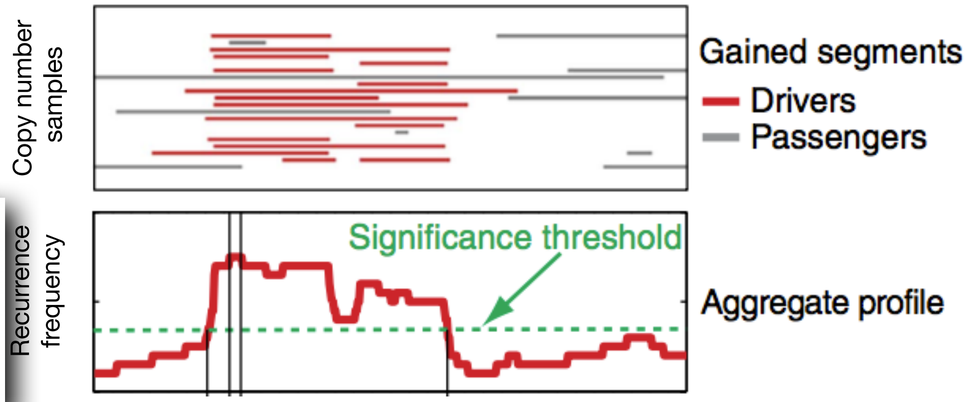


Traditional approach

- method: finding focal signal peaks
- data: point mutation or copy number variation

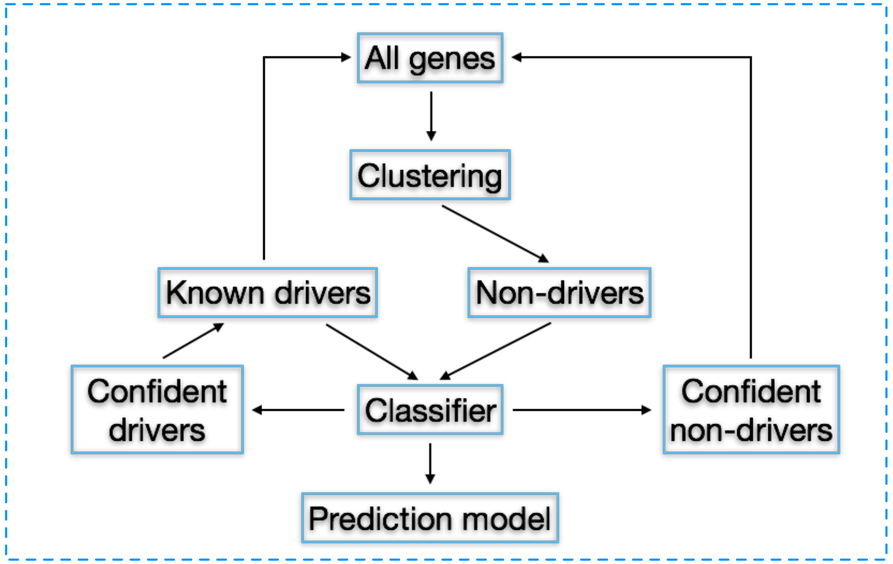
New approach

- method: unsupervised machine learning
- data: point mutation, copy number variation and methylation



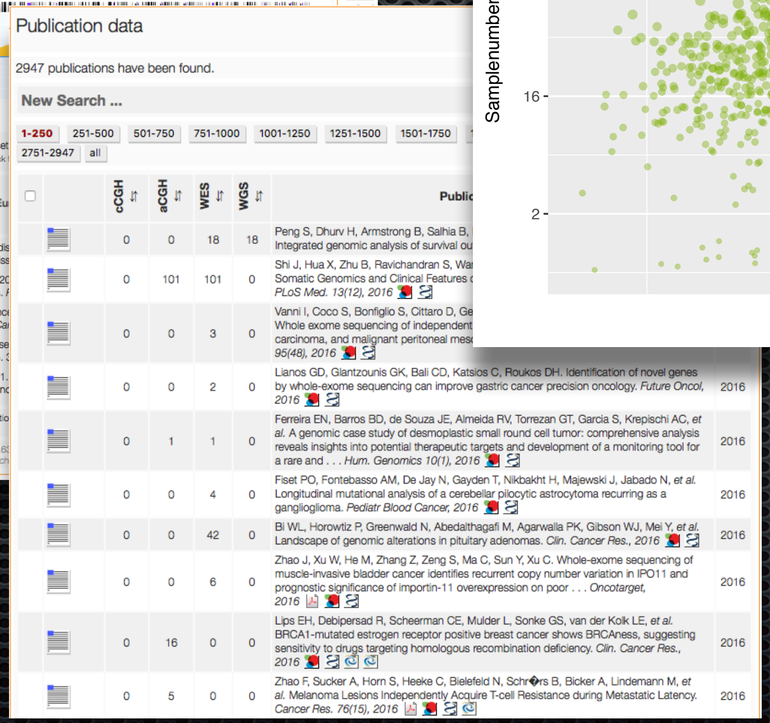
The classical approcah of driver identificaiton

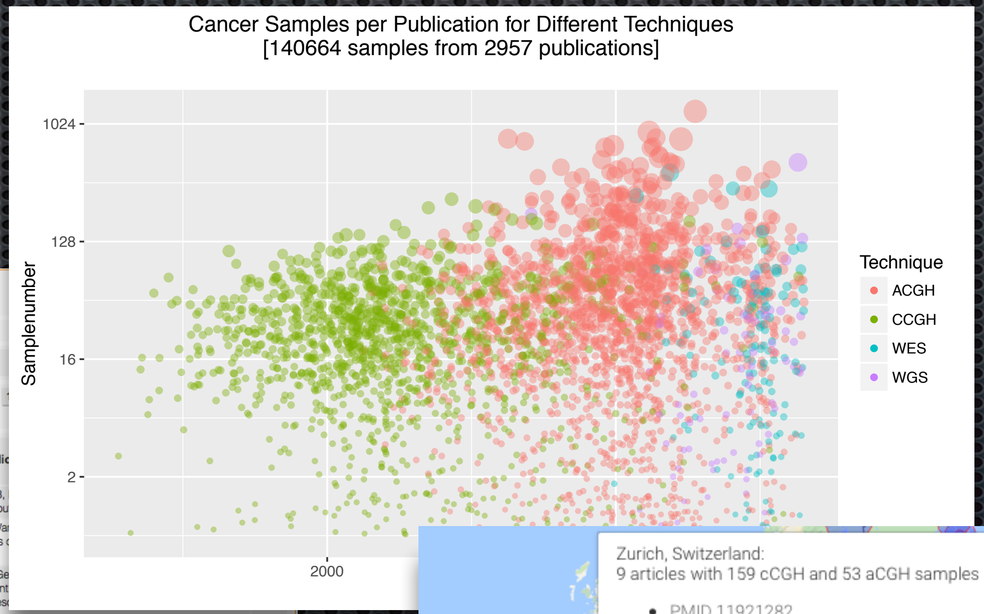
Cited from: Van Dyk, Ewald, et al. *Nature communications* 7 (2016).

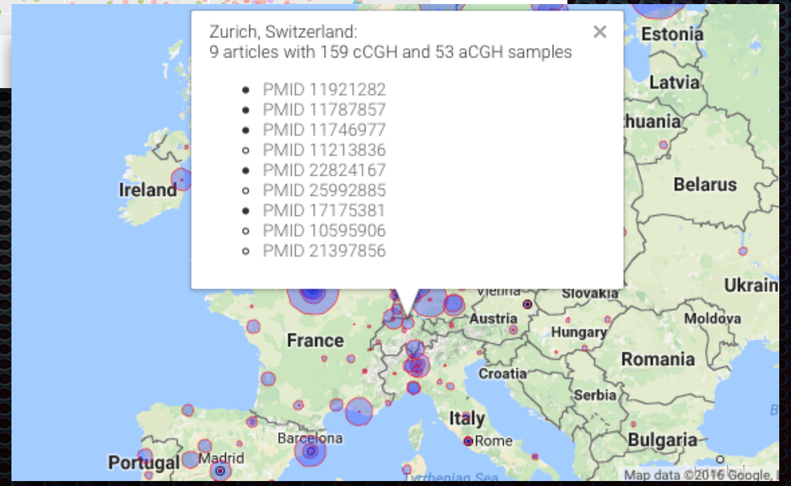


The recursive learning workflow

Progenetix: Cancer Genome Profiles, Article Metrics, Epistemology, **Resource Hub**



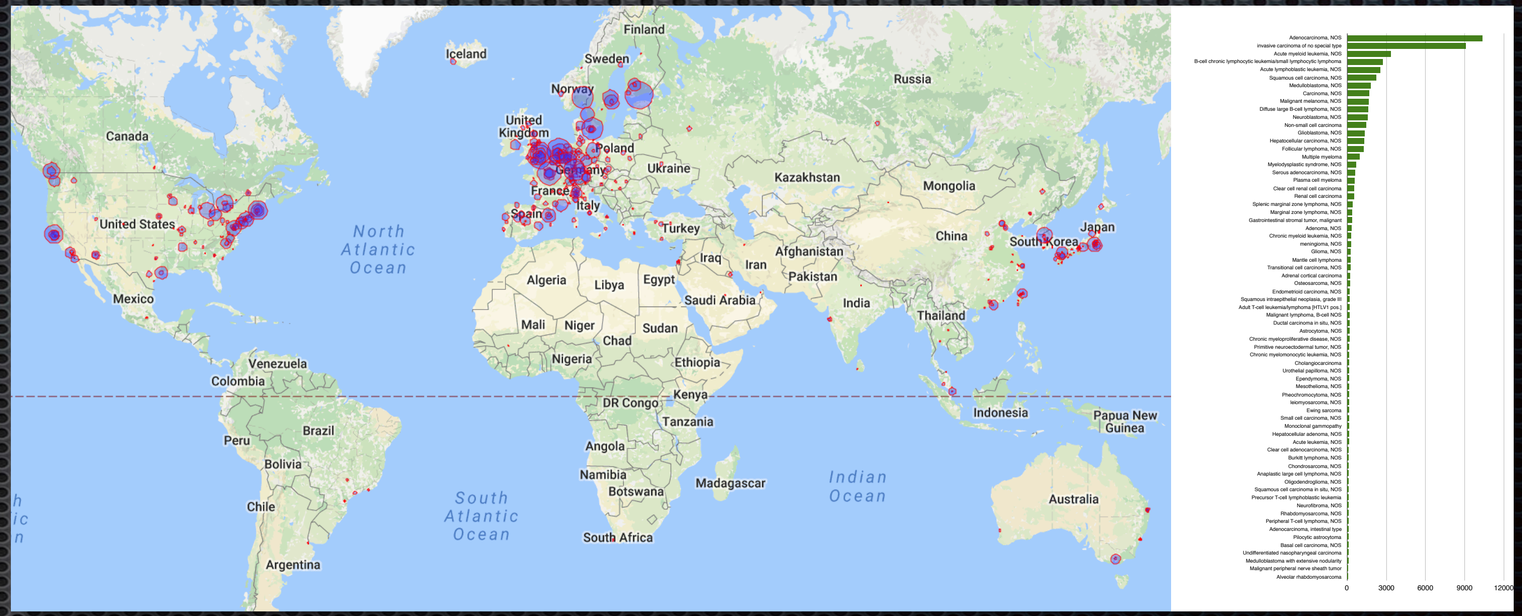




progenetix

Bias in Ascertainment / Background / Environment in Cancer Genome Studies

- the frequency of many genome variants depends on the genetic background
- cancer incidence & type can correlate to environmental factors
- geographic analysis can support interpretation and point to knowledge gaps



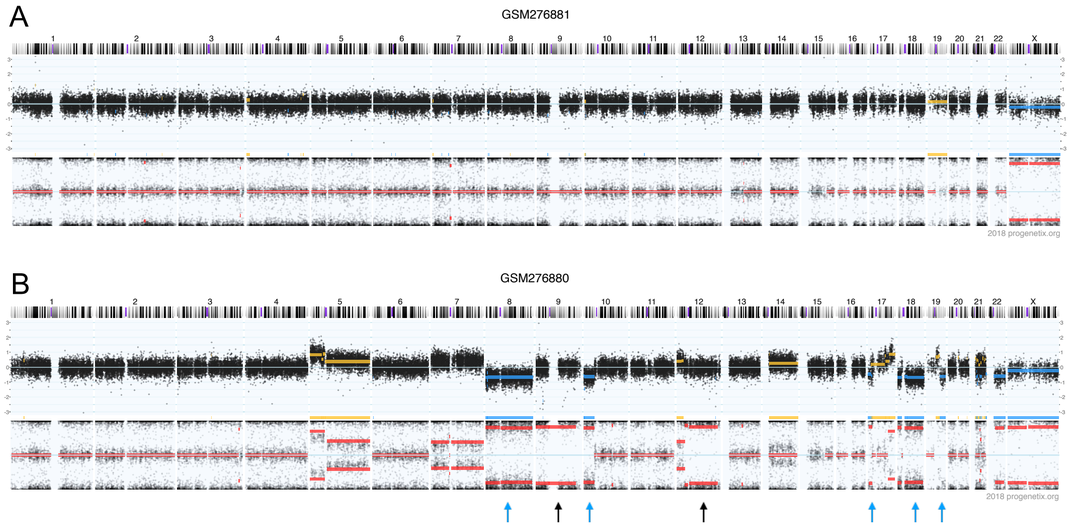
Geographic distribution of >140'000 cancer genome profiles reported in the literature. The numbers are derived from the 2947 publications registered in the Progenetix database.

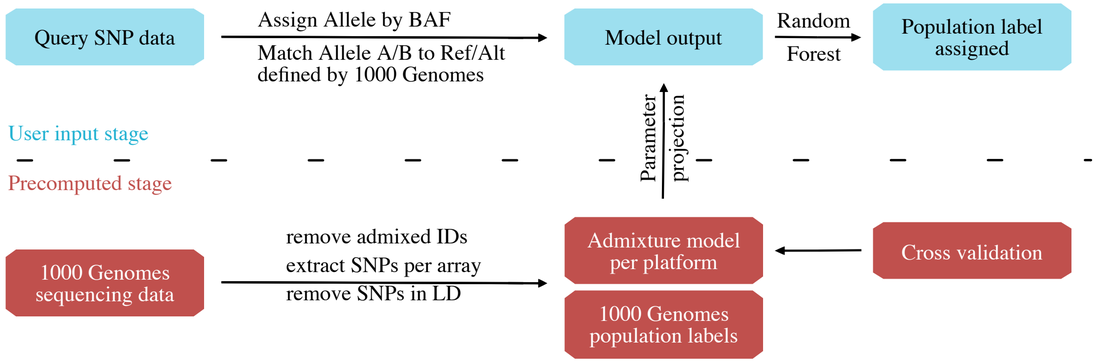
Population stratification in cancer samples based on SNP array data

- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 (or 26) superpopulations: South Asia, Europe, South America, East Asia and Africa.
- SNP positions used in 9 Affymetrix SNP arrays are extracted to train a population admixture model.

Enabling population assignment from cancer genomes with SNP2pop

Qingyao Huang[1,2] and Michael Baudis[1,2]✉





Population stratification in cancer samples based on SNP array data

- Despite extensive somatic mutations of cancer profiling data, consistency between germline and cancer samples reached 97% and 92% for 5 and 26 populations
- Comparison of our benchmarked results with self-reported meta-data estimated a matching rate between 88 % to 92%.
- Ethnicity labels indicated in meta-data are vague compared to the standardized output from our tool

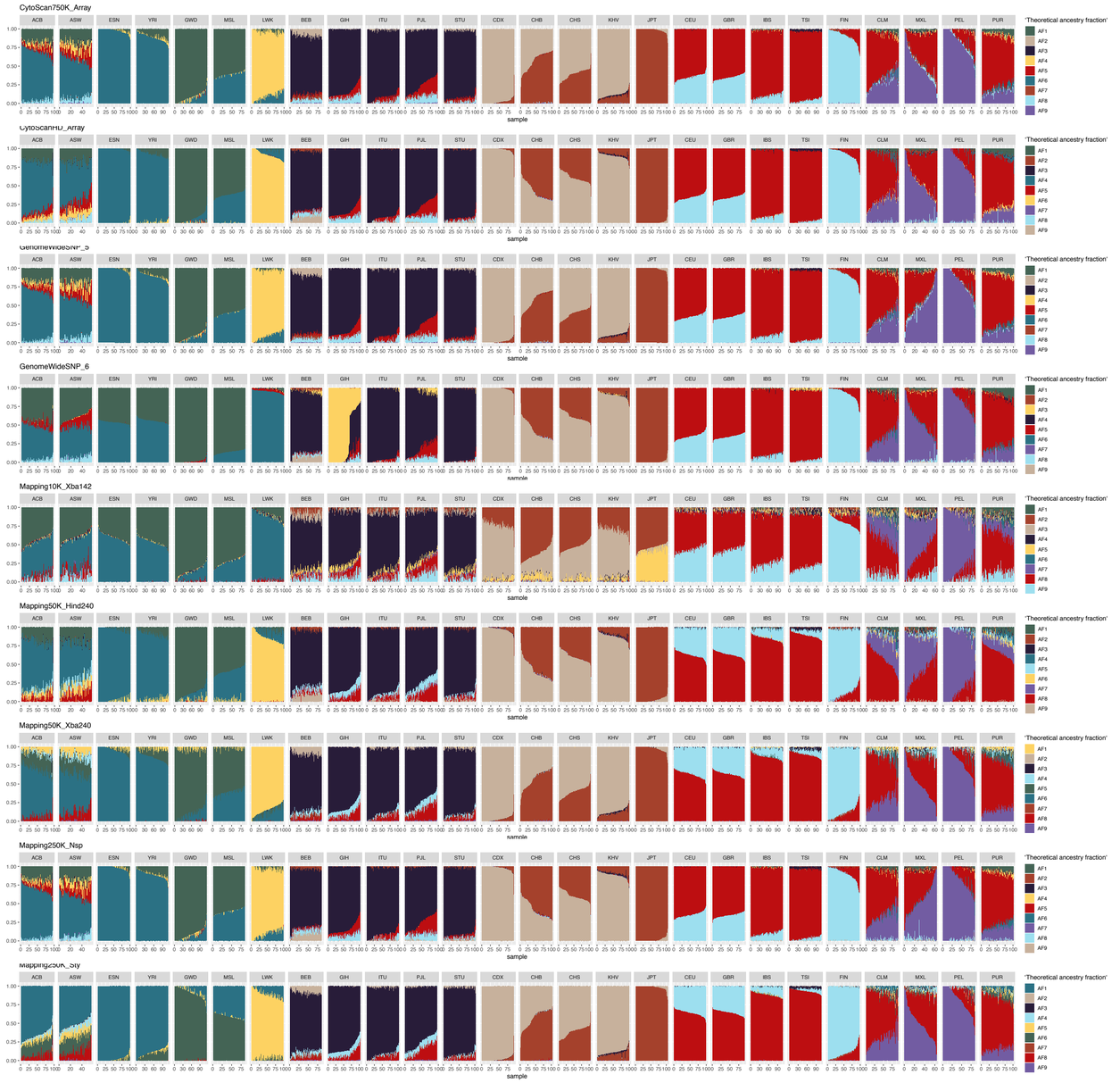
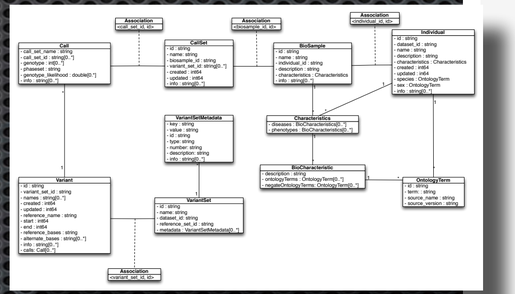


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed neighboring to each other: AFR (1-7), SAS (8-12), EAS (13-17), EUR (18-22), AMR (23-26).

Developing the GA4GH Schemas

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - OntologyTerm objects for biodata
  - implementation w/ ontology services



Driving **Beacon** Development

- Beacon+
  - CNV/CNA as first type of structural variants
  - disease specific queries
  - quantitative reporting

Global Alliance for Genomics & Health

EMBL-EBI

arrayMap

elixir

SIB Swiss Institute of Bioinformatics


```
{
        "_id" : ObjectId("58297ca32ca4591e5a0df054"),
        "id" : "AM_V_1778741",
        "variant_set_id" : "AM_VS_HG18",
        "reference_name" : "10"
        "start" : 579049,
        "end" : 17236099,
        "alternate_bases" : "DUP",
        "reference_bases" : ".",
        "info" : {
          "svlen":16657050,
          "cipos":[
              -1000,
              1000
          ],
          "ciend":[
              -1000,
              1000
          ]
        },
        "calls" : [
                {
                        "genotype" : [
                                ".",
                                "."
                        ],
                        "call_set_id" : "AM_CS_TCGA-61-1917-01A-01D-0648-01",
                        "info" : {
                                "segvalue" : 0.5491
                        }
                }
        ],
        "created" : ISODate("2016-11-14T08:33:58.202Z"),
        "updated" : ISODate("2016-11-14T08:33:58.202Z"),
}
```

Baudisgroup: Wrangling GA4GH Schemas

- object model
  `Individual - Biosample - Callset`
- referencing of ontologies instead of text descriptors
- these are no "real" open ontologies
- data standards use (e.g. ISO)
- fallback to generic object map for unassigned data; this should disappear over time

Global Alliance
for Genomics & Health

```
"id" : "PGX_AM_BS_GSM510730",
"individual_id" : "PGX_IND_GSM510730",
"name" : "PGX_AM_BS_GSM510730",
"description" : "breast carcinoma",
"bio_characteristics" : [
   {
      "description" : "breast carcinoma",
      "ontology_terms" : [
         {
            "term_id" : "ncit:C4017",
            "term_label" : "Ductal Breast Carcinoma"
         },
         {
            "term_id" : "pgx:icdom:8500_3",
            "term_label" : "invasive carcinoma of no special type"
         },
         {
            "term_id" : "pgx:icdot:C50",
            "term_label" : "breast"
         },
         {
      ],
      "negated_ontology_terms" : [ ],
   }
],
"individual_age_at_collection" : "P47Y",
"attributes" : {
   "tnm" : {
      "values" : [
         { "string_value" : "T1N0M0" }
      ]
   },
"location" : {
   "geo_label" : "Oslo, Norway",
   "latitude" : 59.91,
   "longitude" : 10.75,
   "geo_precision" : "city"
},
"external_identifiers" : [
   {
      "database" : "Pubmed",
      "identifier" : "20592421",
      "relation" : "part_of"
   },
],
"updated" : ISODate("2017-03-20T08:37:07.771Z"),
```

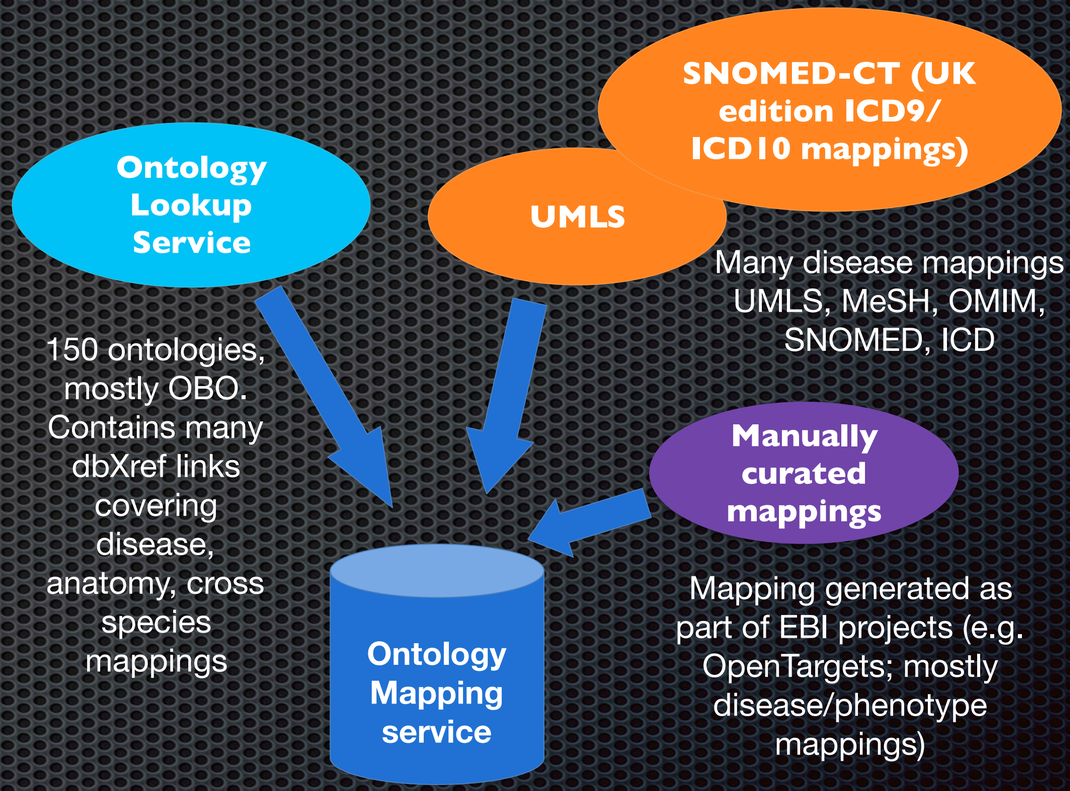
Ontologies need an Einstein to sort them out

University of Zurich[UZH]

ARRAYID	NCIT:CODE	NCIT:TERM	ICD-MORPHOLOGY-CODE	ICD-TOPOGRAPHY-CODE
GSM393758	C2773	Acute Myeloid Leukemia Not Otherwise Specified	9861/3	C42
GSM302275	C2852	Adenocarcinoma	8140/3	C34
GSM918983	C3222	Medulloblastoma	9470/3	C716
GSM551398	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM412374	C3163	Chronic Lymphocytic Leukemia	9823/3	C42
GSM1218276	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM714412	C2852	Adenocarcinoma	8140/3	C569
GSM1109923	C9306	Soft Tissue Sarcoma	8800/3	C499
GSM711848	C2852	Adenocarcinoma	8140/3	C25
GSM746794	C89476		8077/2	C53
GSM1981528	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM2713999	C7949		8500/2	C50
GSM533469	C9349	Plasmacytoma	9731/3	C42

Making Ontologies Work for GA4GH Implementation Studies

- biomedical "metadata" in different resources frequently follows incompatible classification systems
- medical coding systems are driven by different paradigms compared to biological ontologies (e.g. for cross-species comparisons)
- frequently used classifications (ICD, Snomed…) are either not “ontologised” or cannot be referenced in open resources



Federated queries across resources need **curated mappings** of classifications/ ontologies

Ontology Mapping Service OxO

GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

- News
- Participants
- Data Formats
- Data Schemas
- Ontology_te
- Examples, G
- Meeting mi
- Contacts

Related Sites

- GA4GH::Dis
- GA4GH::CLP
- GA4GH::GKS
- SchemaBloc
- ELIXIR Beac
- Phenopacke
- GA4GH
- Beacon+

Tags

Beacon CP admins coordinates

Schemas

Schema elements previously developed as part of various GA4GH efforts had

GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

- News
- Participants
- Data Formats
  - Identifiers and CURIEs
  - Genome Coordinates
  - Dates & Times
- Data Schemas
- Examples, Guides & FAQ
- Meeting minutes
- Contacts

Related Sites

- GA4GH::Discovery
- GA4GH::CLP
- GA4GH::GKS
- SchemaBlocks at Metadata
- ELIXIR Beacon
- Phenopackets
- GA4GH
- Beacon+

Tags

Beacon CP Discovery GA4GH GKS MME admins code contacts contributors coordinates dates developers identifiers leads press times

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

Formats

Schema elements previously developed as part of various GA4GH efforts had been assembled in the SchemaBlocks demonstrator. Those schemas and documentation will be re-implemented in this space.

Additional information about data formats can be found on the GA4GH::Metadata site.

Identifiers and CURIEs

One of the GA4GH conventions is to use CURIEs as (external) identifiers.

mbaudis, 2018-12-24: more ...

Genome Coordinates

This documentation needs to be edited, to represent the GA4GH convention of using "… 0-based, inclusive coordinates".

For now please see

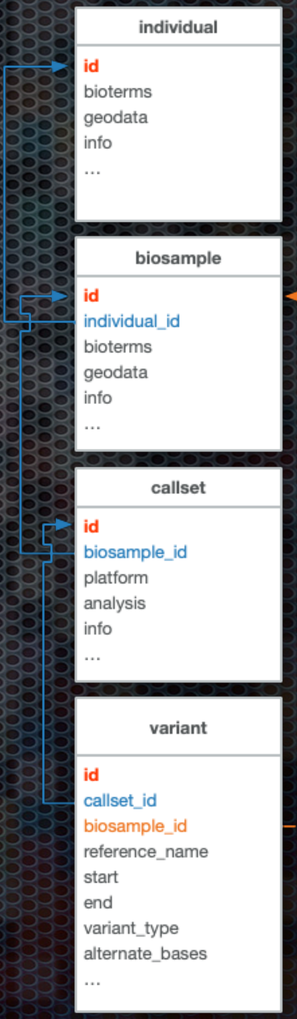
- the documentation of the `Variant` object for the original *GA4GH schema*
- a recent discussion on Github, and the links from there
- a nice explanation of coordinate systems at *Biostars.org* by Obi Griffith

mbaudis, 2018-12-21: more ...

Dates & Times

Date and time formats are specified as ISO8601 compatible strings, both for time points as well as for intervals and durations.

mbaudis, 2018-12-21: more ...



ELIXIR - Towards Biomedical Beacons

Needs & Models Beyond Basic Variant Discovery

Michael Baudis, University of Zurich | **SIB**

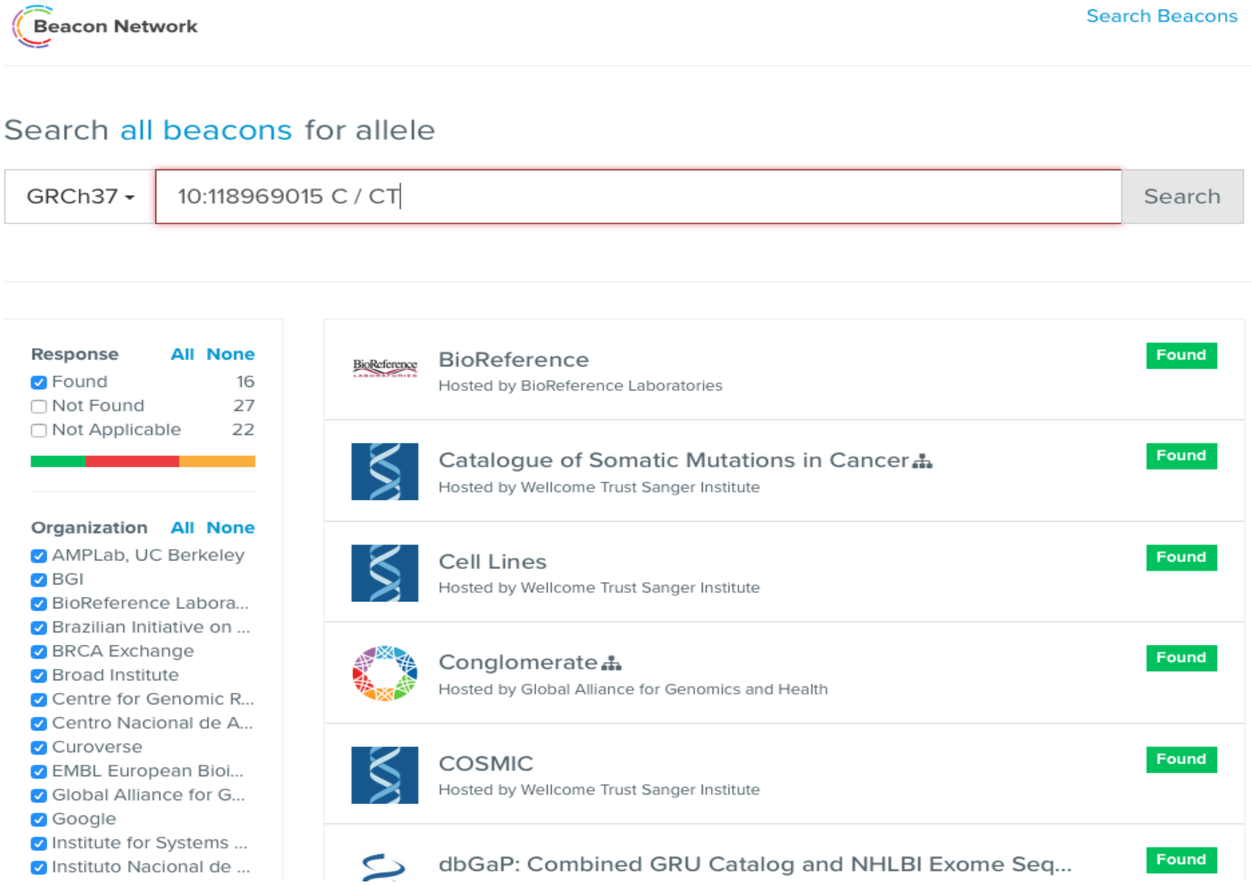
Basel 2019

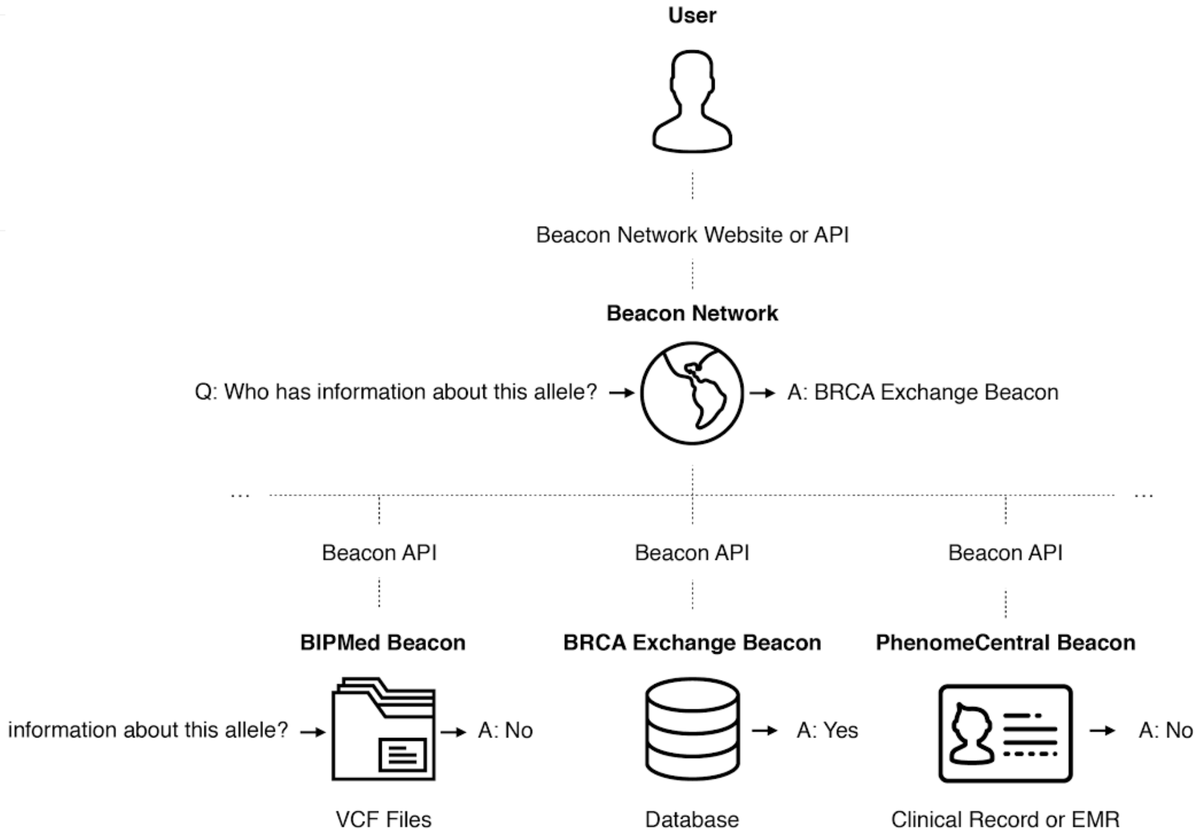
elixir

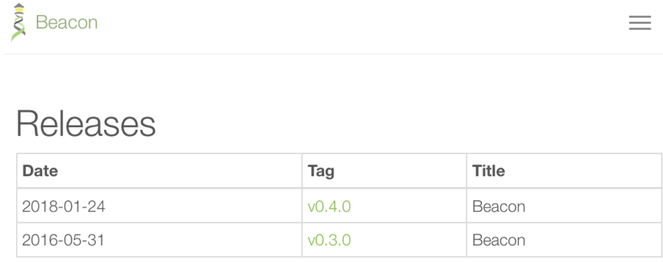
Global Alliance
for Genomics & Health

Beacon Project

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







Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

200+ Datasets

100K+ Individuals

ELIXIR Beacon Project

- Driver project on GA4GH roadmap
- aligns with Discovery Work Stream
- strong impact on GA4GH developments as a concrete, funded project

Beacon *forward*

- ☑ **structural variations** (DUP, DEL) in addition to SNV
- … more structural queries (translocations/fusions…)
- (bio-) **metadata** queries
- ☑ layered authentication system using **ELIXIR AAI**
- quantitative responses
- Beacon queries as entry for **data delivery** (outside Beacon protocol)
- Ubiquitous **deployment** (e.g. throughout ELIXIR network)

Global Alliance for Genomics & Health

beacon.progenetix.org

Beacon Implementations

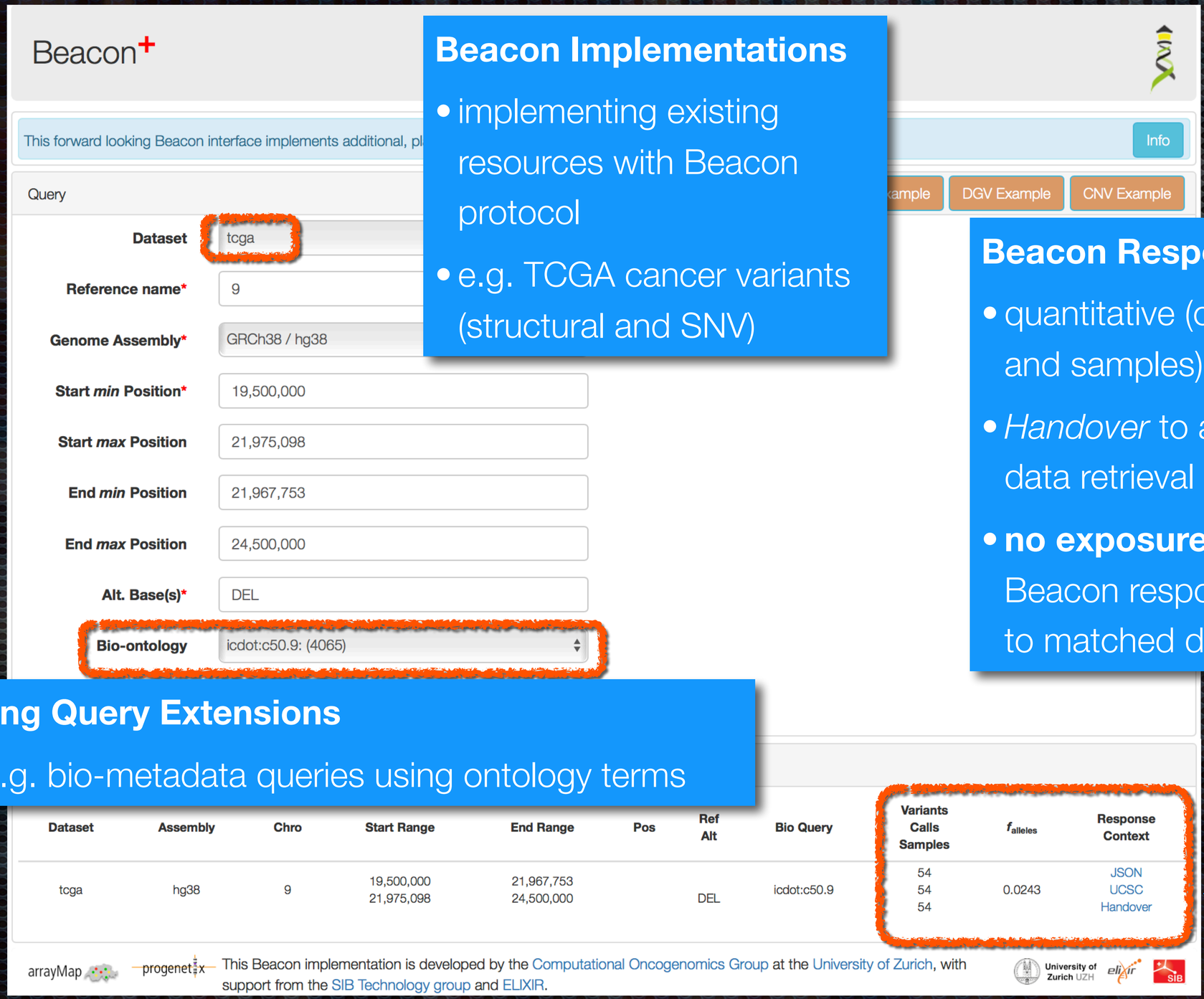
- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

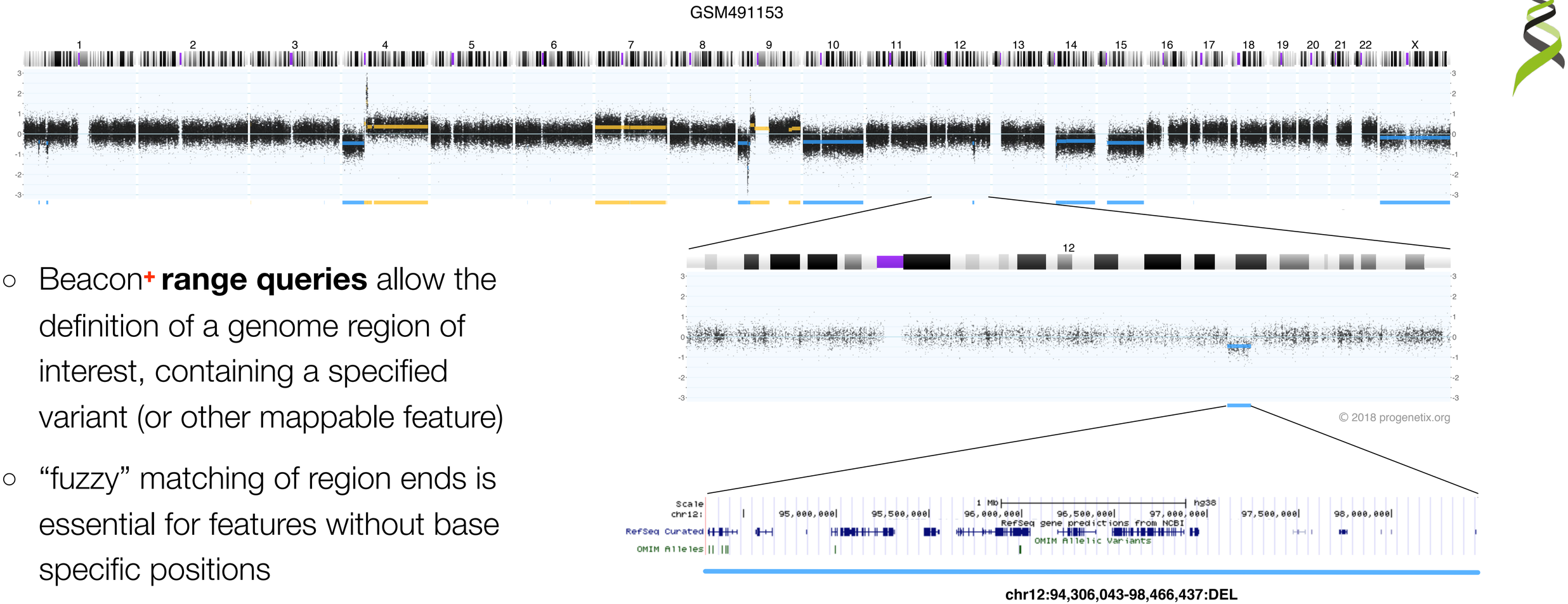
Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

- testing e.g. bio-metadata queries using ontology terms





- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (<DUP>, <DEL>), as are specified in VCF && GA4GH variant schema

chr12:94,306,043-98,466,437:DEL

start_min: 94,000,000
start_max: 94,500,000

variant_type: “**BND**”

reference_name: “9”

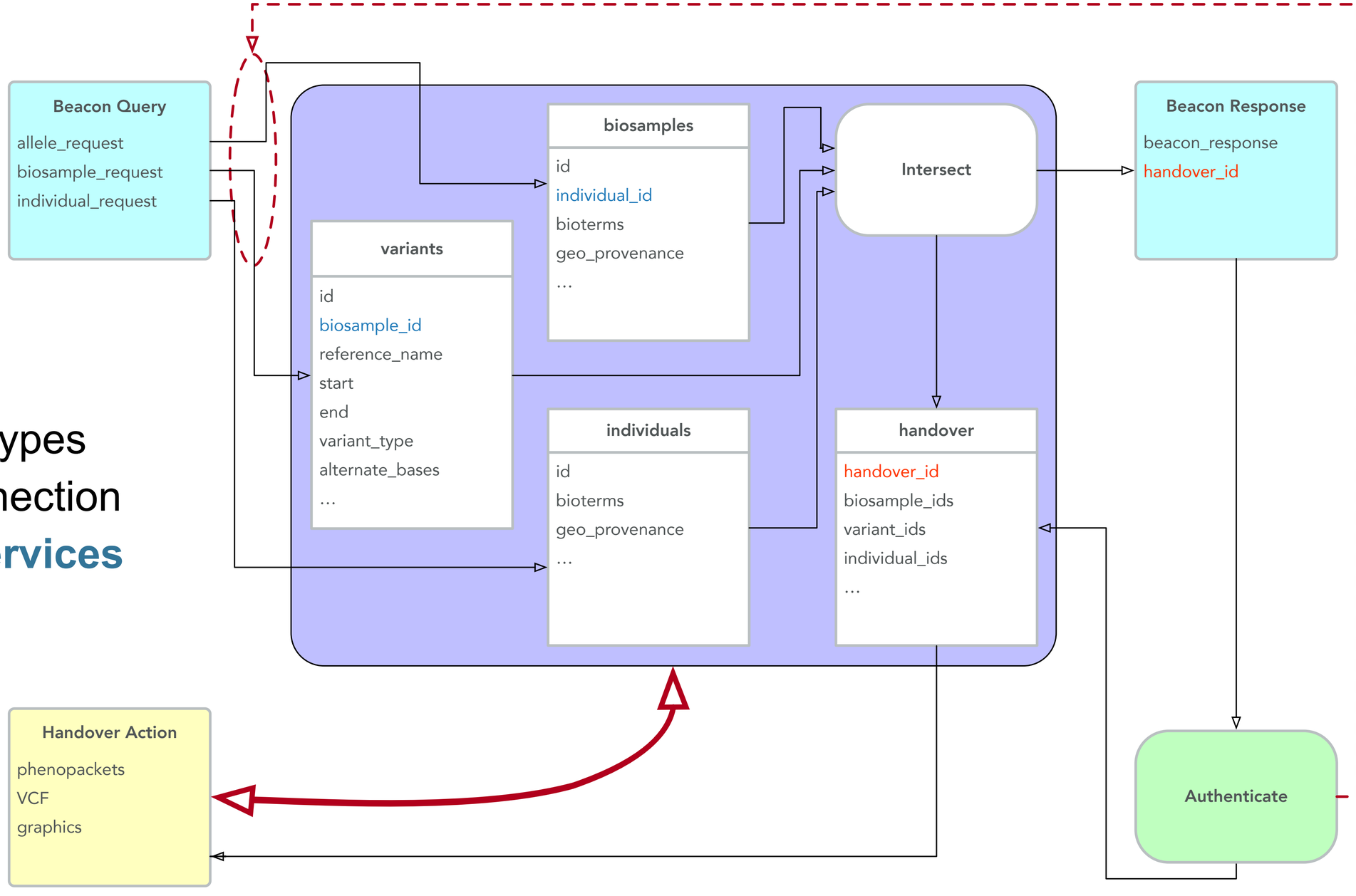
variant_type: “**DEL**”

end_min: 98,200,000
end_max: 98,700,000

variant_type: “**BND**”

Beacon & Handover

Future Beacons to support advanced types of queries and connection of **data delivery services**



Beacon I/O

Handover

Authentication

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

Beacon query => Handover Handle => Authentication => Data Retrieval

Beacon+ example implementation using public somatic variation data

Beacon+

This is an implementation of a Beacon "handover" concept, in which a Beacon query response additionally deli
representation of the query resulte (i.e. callsets, biosamples, metadata ...), which can then be accessed after au
("yes"|"no") or quantitative ("n matches") Beacon response from a data delivery mechanism.

The current implementation exemplifies some possible scenarios:

- providing a histogram of regional gain/loss frequencies (DUP, DEL) for samples with structural variation d
- returning data of the associated callsets which matched the Beacon query (this is for feature demonstrat
- returning the metadata (diagnoses etc.) of the biosamples from which the matching callsets were derivec

This demonstrator does not implement authentication procedures yet; login & password fields can be left empt

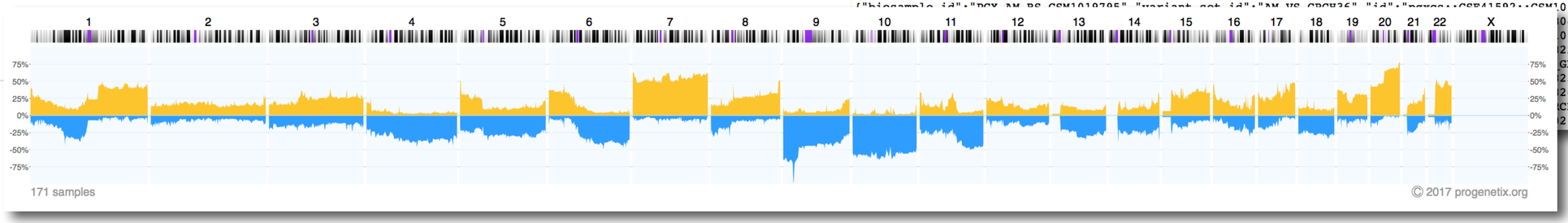
Handover Action

- ✓ Plot DUP/DEL histogram
- Export Callset Data
- Export Biosample Data

Login

Password ••••••••••••••••

Process Data



arrayMap progenetix

Beacon Handover Demonstrator

- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* **response formats not managed by Beacon** protocol - phenopackets, VCF, htsget ...

Global Alliance for Genomics & Health

elixir

Beacon Handover Demonstrator

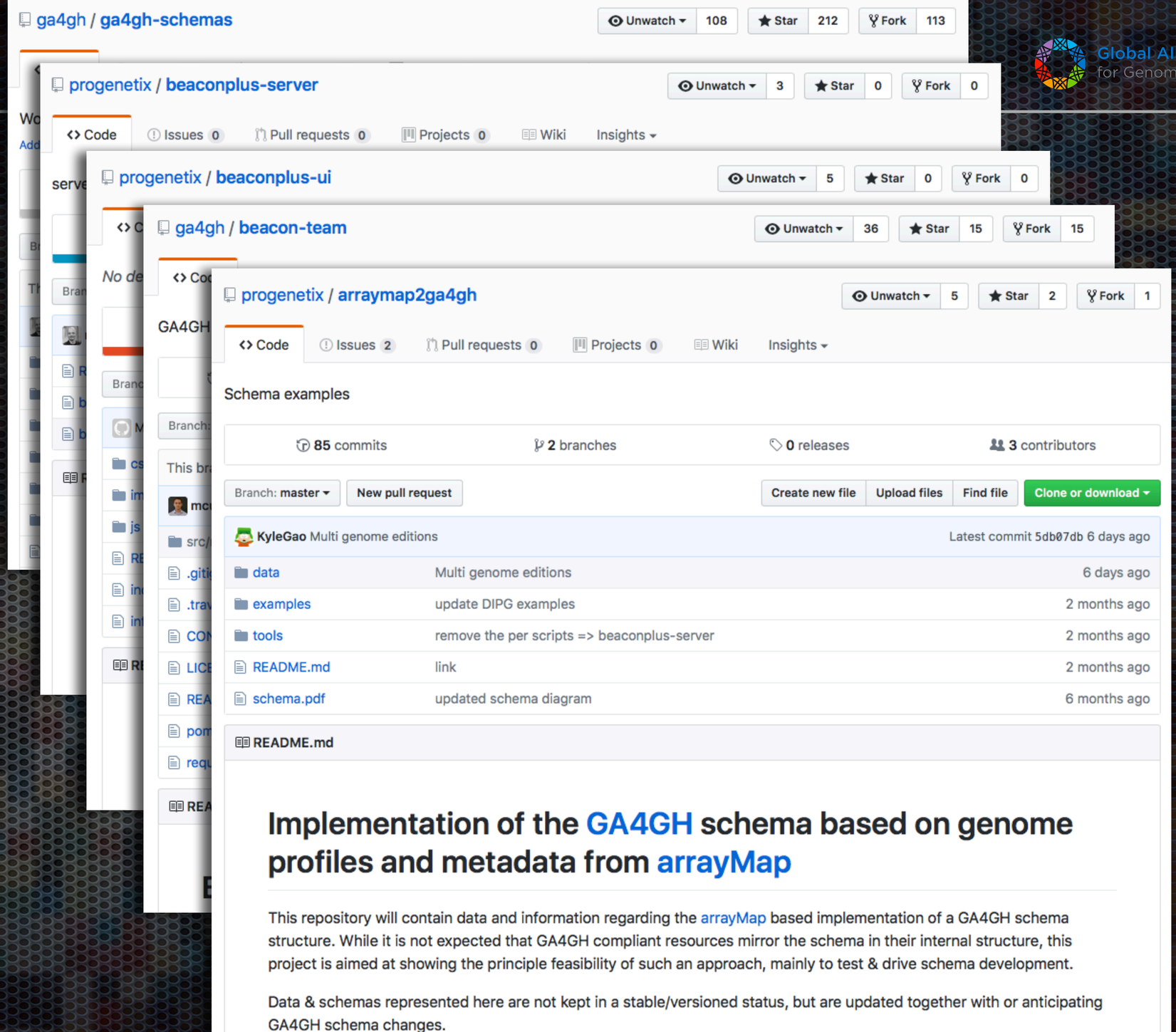
- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* response **outside of Beacon protocol / system**


```
{
  "datasetAlleleResponses": [
    {
      "callCount": 588,
      "variantCount": 588,
      "datasetHandover": [
        {
          "description": "retrieve data of the biosamples matched by the query",
          "handoverType": {
            "id": "pgx:handover:biosamplesdata",
            "label": "Biosamples"
          },
          "url": "https://beacon.progenetix.org/beaconplus-server/
beacondeliver.cgi?do=biosamplesdata&accessid=d50b0a14-1a2c-11e9-af9a-fd65cc50531f"
        },
        {
          "url": "https://beacon.progenetix.org/beaconplus-server/
beacondeliver.cgi?do=cnvhistogram&accessid=d513a3c3-1a2c-11e9-af9a-fa62516f26af",
          "description": "create a CNV histogram from matched callsets",
          "handoverType": {
            "id": "pgx:handover:cnvhistogram",
            "label": "CNV Histogram"
          }
        },
...
```


Global Alliance for Genomics & Health · SIB · University of Zurich UZH · elixir

Check it Out!

- managed, participation driven projects living on Github: ***ga4gh-schemablocks*** & ***ga4gh-beacon***
- test datasets & code available through our ***progenetix*** & ***baudisgroup*** repositories
  - ➡ test
  - ➡ comment
  - ➡ suggest
  - ➡ propose
  - ➡ complain …



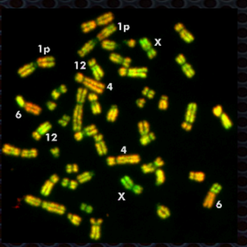
Global Alliance for Genomics & Health

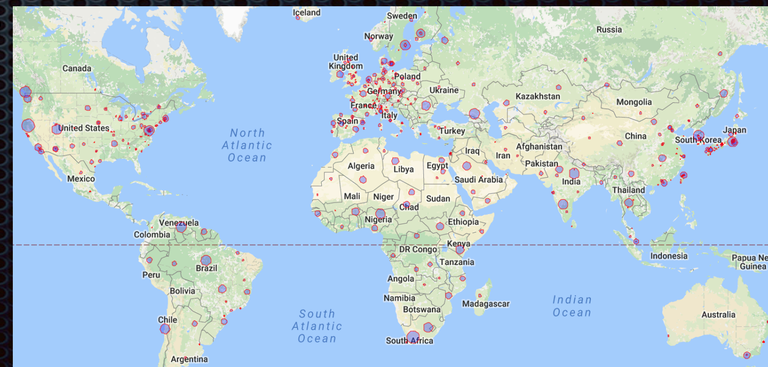
University of Zurich[UZH] SIB elixir

University of Zurich[UZH]

Cancer Genome Data: Where Do We Need To Go

- balancing clinical medicine (**panel** sequencing of **actionable** cancer targets) and the necessary knowledge generation (**complete genomes**, multi"omes", genetic background)
- "genetic **awareness**" and regulatory **security**
- vastly increasing **data curation efforts** to make best use of existing data
- data **sharing frameworks** & **federated** analysis
- **everywhere in the world...**





BAUDISGROUP @ UZH

NI AI
MICHAEL BAUDIS
(HAOYANG CAI)
PAULA CARRIO CORDO
BO GAO
QINGYAO HUANG
SAUMYA GUPTA
(NITIN KUMAR)
(RAHEL PALOOTS)

SIB

AMOS BAIROCH
HEINZ STOCKINGER
DANIEL TEIXEIRA

THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

GA4GH

LARRY BABB
ANTHONY BROOKES
MELANIE COURTOT
MELISSA HAENDEL
MICHAEL MILLER
HELEN PARKINSON
GUNNAR RÄTSCH
ANDY YATES

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
GARY SAUNDERS
ILKKA LAPPALAINEN
S. DE LA TORRE PERNAS
SERENA SCOLLEN
JUHA TÖRNROOS

University of Zurich[UZH]

SIB

elixir

Global Alliance for Genomics & Health

University of Zurich[UZH]

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org
progenetix.org
info.baudisgroup.org
sib.swiss/baudis-michael
imls.uzh.ch/en/research/baudis
beacon-project.io
schemablocks.org

Global Alliance
for Genomics & Health