Genome data access

Somatic CNV databases as resources for cancer research && GA4GH schema development

EMBL-EBI elixir SIB University of Zurich UZH Global Alliance for Genomics & Health

1992	UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386 dkfz.	Heidelberg	Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics
2001	Stanford University	Stanford	Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website
2003	UF UNIVERSITY of FLORIDA	Gainesville	Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis
2006	RWTH AACHEN UNIVERSITY	Aachen	Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities
2007	SIB Swiss Institute of Bioinformatics; Universität Zürich[UZH]	Zürich	Professor of bioinformatics @ IMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *arraymap* online resource | GA4GH | SPHN

Cancer Genome Data & Beyond

Contributing our share

- 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!

arrayMap

progenetix

Global Alliance for Genomics & Health

SAMW ASSM

akademien der wissenschaften schweiz

SIB Swiss Institute of Bioinformatics

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 genome information and individually targeted therapies.

Genome Sequencing

whole genome sequencing (WGS)

100%

all DNA
(3.1 billion base pairs)

exome sequencing

~1%

protein-coding DNA only
(~31 million base pairs)

What does it cost to sequence a genome?

	Human Genome Project (HGP): 1991-2003	today: 2017	today: 2017
cost:	$2.7 billion	~$1,500	~$530
time:	12+ years	< 2 days	~3 days

Ryan L. Collins & Brad Wierbowski (2017)

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

* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

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)

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

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

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	currently data from >16'000 samples
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

Global Alliance for Genomics & Health

Genome Data Access in Cancer

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 articles assessed for the progenetix.org cancer genome data resource (M. Baudis, 2001-2016)

Genome-wide multi"*omics*" data generation for understanding tumor biology

- the International Cancer Genome Consortium (ICGC) as leading example of deep analysis of multiple cancer entities
- international collbaoration of leadinc research centers for each of ~20 tumor types
- limitations:
  - focus on prominant cancer types w/ limited representation of rare entities
  - data access policies influenced by national regulations and legal frameworks
  - technical heterogeneity

GA4GH to solve accessibility…

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

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

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

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

Global Alliance for Genomics & Health

This repository Search Pull requests Issues Gist

ga4gh / schemas

Unwatch 115 Star 196 Fork 110

Code Issues 152 Pull requests 29 Projects 1 Wiki Pulse Graphs

Work on data models and APIs for Genomic data. http://ga4gh.org/#/api

1,102 commits 17 branches 16 releases 46 contributors Apache-2.0

Branch: metadata-integ... New pull request Create new file Upload files Find file Clone or download

This branch is 15 commits ahead, 3 commits behind master. Pull request Compare

mbaudis Merge branch 'master' into metadata-integration Latest commit 077c2c7 2 days ago

doc	Merge branch 'master' into metadata-integration	2 days ago
python	Add constraints file	2 days ago
scripts	Utilize new common methods in schemas	2 days ago
src/main/proto	Merge branch 'master' into metadata-integration	13 days ago
tests	Utilize new common methods in schemas	2 days ago
tools	Merge branch 'master' into metadata-integration	13 days ago
.gitignore	Remove protoc call from install path (#781)	7 days ago
.travis.yml	Add constraints file	2 days ago
CONTRIBUTING.rst	Convert Avro -> proto3.	10 months ago

The State of the Schema, Feb 2017

Global Alliance
for Genomics & Health

Global Alliance
for Genomics & Health

R+E

Security

Dissemination

Real world projects

Rare Disease

Cancer

Traits + Chronic D

Basic Biology

Workstreams

Discovery

Large Scale Genomics

Access + Authentication

Cloud

Reference

Phenotype (includes eHealth bridge)

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.

The Cancer Gene Trust
proposes to aggregate somatic cancer mutation data and some clinical data in order to improve the genomic landscape of actionability in some cancers and to enable greater personalized clinical care for individuals with rare cancer mutations.

BRCA CHALLENGE

Matchmaker Exchange

Cancer Gene Trust

Beacon Project

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

Global Alliance
for Genomics & Health

Genome ***Beacons*** Compromise Security?

Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

Stanford researchers identify potential security hole in genomic data-sharing network

Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29
2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.
Science photo/Shutterstock

Work by a pair of researchers at the Stanford University School of Medicine makes that genomic data more secure. Suyash Shringarpure, PhD, a postdoctoral scholar in genetics, and Carlos Bustamante, PhD, a professor of genetics, have demonstrated a technique for hacking a network of global genomic databases and how to prevent it. They are working with investigators from the Global Alliance for Genomics and Health on implementing preventive measures.

The work, published Oct. 29 in *The American Journal of Human Genetics,* also bears importantly on the larger question of how to analyze mixtures of genomes, such as those from different people at a crime scene.

Cancer Genome Data & Beyond

Contributing our share

- 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!

arrayMap

progenetix

Global Alliance for Genomics & Health

SAMW ASSM

akademien der wissenschaften schweiz

SIB Swiss Institute of Bioinformatics

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

ARRAYMAP DATA “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

Results of Liftover

- Convert from hg18 to hg19 and GRCh38
- Processed 122,788 files, 26,164,205 segments and 28,941,899,671 probes in total
- single run took more than a week
- parallel run w/ 4 processes took less than 3 days
- Network & file server are the bottleneck
- Reduced data loss (10% to 0.1%)

Distribution of remapped and unmapped

Distribution of unmapped probes

Reason of unmapped segments

arrayMap

Resource for copy number variation data in cancer

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)

Subsets	Samples	Observations	Frequency
9590/3: 9590/3: Malignant lymphoma, NOS	6	5	0.833
9052/3: 9052/3: Epithelioid mesothelioma, malignant	22	14	0.636
9050/3: 9050/3: Mesothelioma, malignant	87	55	0.632
9729/3: 9729/3: Precursor T-cell lymphoblastic lymphoma	30	14	0.467
8130/1: 8130/1: Papillary transitional cell neoplasm of low malignant potential	38	12	0.316
8250/3: 8250/3: Bronchiolo-alveolar adenocarcinoma, NOS	23	7	0.304
9801/3: 9801/3: Acute leukemia, NOS	10	3	0.300
9440/3: 9440/3: Glioblastoma, NOS	2047	607	0.297
8560/3: 8560/3: Adenosquamous carcinoma	21	6	0.286
8072/3: 8072/3: Squamous cell carcinoma, large cell, nonkeratinizing, NOS	96	27	0.281
9827/3: 9827/3: Adult T-cell leukemia/lymphoma (HTLV-1 positive)	179	49	0.274
9837/3: 9837/3: Precursor T-cell lymphoblastic leukemia	233	59	0.253
8500/2: 8500/2: Intraductal carcinoma, noninfiltrating, NOS	147	36	0.245
8430/3: 8430/3: Mucoepidermoid carcinoma	21	5	0.238
8811/3: 8811/3: Fibromyxosarcoma	43	10	0.233
8160/3: 8160/3: Cholangiocarcinoma	39	9	0.231
8012/3: 8012/3: Large cell carcinoma, NOS	48	11	0.229

arrayMap

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

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.

progenetix

Population stratification in cancer samples based on SNP array data

- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 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.

Developing the GA4GH **Metadata** Schema

- 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"),
}
```

GA4GH DWG Metadata Task Team
Wrangling the Schema

- 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


```
"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"),
```


Global Alliance
for Genomics & Health

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
GSM121826	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM714412	C2852	Adenocarcinoma	8140/3	C569
GSM110923	C9306	Soft Tissue Sarcoma	8800/3	C499
GSM711848	C2852	Adenocarcinoma	8140/3	C25
GSM746794	C89476		8077/2	C53
GSM198528	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM271399	C7949		8500/2	C50
GSM533469	C9349	Plasmacytoma	9731/3	C42

Working towards ontologies w/ arrayMap: Mapping >55'000 samples from ICD-O to NCIt neoplasm core

example_dx	ICDMORPHOLOGY	ICDOM	ICDTOPOGRAPHY	ICDOT	NCIT:CODE
malignant melanoma [metastatic cell line MaMel19]	Malignant melanoma NOS	8720/3	skin	C44	C3224
malignant melanoma [vagina]	Malignant melanoma NOS	8720/3	vagina and labia	C510	C3224
malignant melanoma [uvea metastasized]	Malignant melanoma NOS	8720/3	retina	C692	C3224
meningioma	Meningioma NOS	9530/0	meninges cerebral spinal	C700	C3230
mesothelioma	Mesothelioma NOS	9050/3	lung and bronchus	C34	C3234
pleural mesothelioma	Mesothelioma NOS	9050/3	pleura	C384	C3234
mesothelioma	Mesothelioma NOS	9050/3	connective and soft tissue NOS	C499	C3234
multiple myeloma	Plasma cell myeloma	9732/3	hematopoietic and reticuloendothelial syste	C42	C3242
Mycosis fungoides	Mycosis fungoides	9700/3	skin	C44	C3246
Myelodysplastic syndrome	Myelodysplastic syndrome NOS	9989/3	hematopoietic and reticuloendothelial syste	C42	C3247
Acute myeloblastic leukemia with maturation [FAB M2]	Acute myeloblastic leukemia with maturation [FAB M2]	9874/3	hematopoietic and reticuloendothelial syste	C42	C3250
neuroblastoma	Neuroblastoma NOS	9500/3	peripheral nervs incl. autonomous	C47	C3270
Cerebral neuroblastoma [cerebral region midline frontal lobe]	Neuroblastoma NOS	9500/3	cerebrum	C710	C3270
neuroblastoma [adrenal gland cell line]	Neuroblastoma NOS	9500/3	adrenal gland	C76	C3270
Cutaneous neurofibroma	Neurofibroma NOS	9540/0	skin	C44	C3272
Plexiform neurofibroma	Neurofibroma NOS	9540/0	Nervous system NOS	C729	C3272
Oligodendroglioma [Supratentorial Frontal Lobe]	Oligodendroglioma NOS	9450/3	cerebrum	C710	C3288
oilgodendroglioma	Oligodendroglioma NOS	9450/3	Brain NOS	C719	C3288
oligodendroglioma	Oligodendroglioma NOS	9450/3	brain nos	c719	C3288
Paraganglioma	Paraganglioma NOS	8680/1	Nervous system NOS	C729	C3308
paraganglioma	paraganglioma NOS	8680/1	adrenal cortex	C740	C3308
Pheochromocytoma	Pheochromocytoma NOS	8700/0	adrenal cortex	C740	C3326
polycythemia vera	Polycythemia vera	9950/3	hematopoietic and reticuloendothelial syste	C42	C3336
pediatric rhabdomyosarcoma	Rhabdomyosarcoma NOS	8900/3	connective and soft tissue NOS	C499	C3359
Sezary syndrome	Sezary syndrome	9701/3	hematopoietic and reticuloendothelial syste	C42	C3366
sezary syndrome	sezary syndrome	9701/3	skin	C44	C3366
Synovial sarcoma	Synovial sarcoma NOS	9040/3	connective and soft tissue NOS	C499	C3400
essential thrombocythemia	Essential thrombocythemia	9962/3	hematopoietic and reticuloendothelial syste	C42	C3407
carcinosarcoma	Carcinosarcoma NOS	8980/3	connective and soft tissue NOS	C499	C34448
Carcinosarcoma [breast cell line HS578T]	Carcinosarcoma NOS	8980/3	breast	C50	C34448
acute monocytic leukemia	Monocytic leukemia NOS	9860/3	hematopoietic and reticuloendothelial syste	C42	C3171
leiomyoblastoma	Epithelioid leiomyoma	8891/0	kidney	C649	C3157
colon mucosa [low grade dysplastic tumor; myhmut]	atypical adenoma	8140/1	large intestine excl. rectum and rectosigmoi	C189	C7559

Variants from arrayMap

- name here could be e.g. an rsid

- calls are embedded in the variant set
- alternative option?

- one call here
- "info" provides an intensity measurement
- could e.g. provide QC, copy number count …

- Housekeeping needed?


```
{
    "_id" : ObjectId("589dfc4409d374e4f3763266"),
    "id" : "AM_V_622129",
    "name" : null,
    "variant_set_id" : "AM_VS_HG18",
    "reference_name" : "7",
    "start" : 157559617
    "reference_bases" : ".",
    "end" : 158812247,
    "alternate_bases" : "DUP",
    "info" : {},
    "calls" : [
        {
            "call_set_id" : "AM_CS_GSM322223",
            "genotype" : [
                ".",
                "."
            ],
            "info" : {
                "segvalue" : 0.2698
            }
        }
    ],
    "updated" : ISODate("2017-02-10T17:15:02.380Z"),
    "created" : ISODate("2017-02-10T17:15:02.380Z"),
}
```

svlen
(200)
start
(100)
variant_type
(DEL)
end
(300)
Minimal exact match
q_start
[100]
q_type
“DEL”
q_end
[300]
Match using query ranges
q_start
[80,120]
q_type
“DEL”
q_end
[260,360]
Match using query ranges “left exact right open”
q_start
[100]
q_type
“DEL”
q_end
[260,360]
Match using query ranges “at least one base in interval affected”
q_start
[100,300]
q_type
“DEL”
q_end
[100,300]

BeaconCNVrequest - First of Beacon+

- First of many (?) request types to propel Beacon from a *social experiment* into a genomics and health data dicovery tool
- imntroduction of a new record/query object does not interfere with current Beacon implementations
- Use of ranges for start, end as a **prototype** for more general introduction of range matches (e.g. “all TAA in a range corresponding to the CDR of gene ABC”)

```
message BeaconAlleleRequest {
  // Reference name (chromosome).
  //
  // Accepted values: 1-22, X, Y.
  // Required.
  string reference_name = 1;

  // Position, genomic locus (0-based).
  //
  // start only:
  //   - for single positions, e.g. the start of a specified sequence alteration
  //     where the size is given through the specified alternate_bases
  //   - typical use are queries for SNV and small InDels
  //   - the use of "start" without an "end" parameter requires the use of
  //     "reference_bases"
  //
  // start and end:
  //   - special use case for exactly determined structural changes
  //
  // start_min + start_max + end_min + end_max
  //   - for querying imprecise positions (e.g. identifying all structural
  //     variants starting anywhere between start_min <-> start_max, and ending
  //     anywhere between end_min <-> end_max
  //   - single or douple sided precise matches can be achieved by setting
  //     start_min = start_max XOR end_min = end_max
  //
  int64 start = 2;

  int64 start_min = 3;

  int64 start_max = 4;

  int64 end = 5;

  int64 end_min = 6;

  int64 end_max = 7;

  // Reference bases for this variant (starting from `start`).
```

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap data
- structural variations
- quantitative queries
- metadata
- current version uses GA4GH schema compatible, non-SQL database backend (MongoDB)

Beacon+

This experimental Beacon interface currently implements **structural variant requests** (i.e. DUP, DEL; see Beacon Github PR #94), in addition to standard Beacon SNV calls.

Code for the project is accessible through the Progenetix Github repository:

- frontend (JS & HTML) and server code (Perl CGI)
- example datasets (MongoDB variant, biosample ... collections)

Query

Dataset

DIPG (CNV + selected SNV)

Reference name*

chromosome name, e.g: 9

Variant type*

SNV / indel

Position*

position for exact allelic variant, e.g. 7577121

Ref. Base(s)*

the reference base for the allele, e.g. G

Alt. Base(s)*

the SNV variant for the allele, e.g. A

Beacon Query

Beacon+

This experimental Beacon interface currently implements **structural variant requests** (i.e. DUP, DEL; see Beacon Github PR #94), in addition to standard Beacon SNV calls.

Code for the project is accessible through the Progenetix Github repository:

- frontend (JS & HTML) and server code (Perl CGI)
- example datasets (MongoDB variant, biosample ... collections)

Query

Dataset

arrayMap (CNV only)

Reference name*

chromosome name, e.g: 9

Variant type*

DEL (Deletion)

Start *min* Position*

begin of match start interval, e.g: 20000000

Start *max* Position

end of match start interval, e.g: 21984490

End *min* Position

begin of match end interval, e.g: 21984490

End *max* Position

end of match end interval, e.g: 25000000

Beacon Query

SIB elixir Global Alliance for Genomics & Health arrayMap

```
{
   "allele_request" : {
      "$and" : [
         { "reference_name" : “17" },
         { "variant_type" : “DEL" },
         { "start" : { "$gte" : 5000000 } },
         { "start" : { "$lte" : 7578000 } },
         { "end" : { "$gte" : 7577000 } },
         { "end" : { "$lte" : 9000000 } }
      ]
   },
   "api_version" : "0.4",
   "beacon_id" : "org.progenetix:progenetix-beacon",
   "dataset_allele_responses" : [
      {
         "call_count" : "162",
         "dataset_id" : "arraymap",
         "error" : null,
         "exists" : true,
         "external_url" : "http://arraymap.org",
         "frequency" : 0.0165,
         "info" : {
            "bs_direct" : null,
            "description" : "The query was against database \"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 162 / 9434 matched callsets for
893279 variants. Out of 9818 biosamples in the database, 9818 matched the biosample query; of those, 162 had the variant.",
            "ontology_selection" : [
               "NCIT"
            ],
            "phenotype_response" : []
         },
         "note" : "",
         "sample_count" : "162",
         "variant_count" : 144
      }
   ],
   "exists" : true,
   "info" : {
      "biosample_Request" : {},
      "query_string" :
"dataset_id=arraymap&variants.reference_name=chr17&variants.variant_type=DEL&variants.start_max=5000000&variants.start_min=7578000&variants.end_min=7577000&v
ariants.end_max=9000000",
      "version" : "Beacon+ implementation based on a development branch of the beacon-team project: https://github.com/ga4gh/beacon-team/pull/94"
   },
   "url" : "http://progenetix.org/beacon/info/"
}
```

Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap data
- structural variations
- quantitative queries
- metadata
- current version uses GA4GH schema compatible, non-SQL database backend (MongoDB)

Beacon+

This experimental Beacon interface currently implements **structural variant requests** (i.e. DUP, DEL; see Beacon Github PR #94), in addition to standard Beacon SNV calls.

Code for the project is accessible through the Progenetix Github repository:

- frontend (JS & HTML) and server code (Perl CGI)
- example datasets (MongoDB variant, biosample ... collections)

Query

Dataset DIPG (CNV + selected SNV)

Reference name* 17

Variant type* SNV / indel

Position* 7577121

Ref. Base(s)* G

Alt. Base(s)* A

Beacon Query

Response

Dataset	Reference name	Variant Type	Start Min	Start Max	End Min	End Max	Position	Ref Bases	Alt Bases	Call Count	Frequency	Query
arraymap	17	DEL	5000000	7578000	7577000	9000000				162	0.0165	show JSON
dipg	17	SNV					7577121	G	A	21	0.0197	show JSON

SIB elixir Global Alliance for Genomics & Health arrayMap

```
{
   "beaconId" : "dipg-ga4gh",
   "callCount" : "144",
   "info" : "The query was against database \"dipg_ga4gh\", variant collection
\"variants_cnv_grch36\". 144 / 834 matched callsets for 199633 variants. Out of 1067
biosamples in the database, 740 matched the biosample query; of those, 116 had the
variant.",
   "error" : null,
   "alleleRequest" : {
      "$and" : [
         { "reference_name" : “9" },
         { "variant_type" : “DEL" },
         { "start" : { "$gte" : 19000000 } },
         { "start" : { "$lte" : 21984490 } },
         { "end" : { "$gte" : 21984490 } },
         { "end" : { "$lte" : 25000000 } }
      ]
   },
   "frequency" : 0.1087,
   "datasetAlleleResponses" : 116,
   "biosampleRequest" : {
      "bio_characteristics.ontology_terms.term_id" : "PGX:ICDOT:C719"
   },
   "sampleCount" : 116
}
```

Implementing real-world datasets for federated access using GA4GH schema specifications: pHGG

Mackay A, Jones C, Baudis M and many, many others:
Integrated molecular meta-analysis of 1000 paediatric high grade and diffuse intrinsic pontine glioma (2017, submitted)

```
{
    "_id" : ObjectId("591eb73709d374442fcc6b2a"),
    "individual_id" : "DIPG_IND_0809",
    "id" : "DIPG_BS_0809",
    "name" : "pHGG_META_0809",
    "description" : “glioma, paediatric, high grade",
    "individual_age_at_collection" : {
        "age_class" : {
            "term" : "Adult onset",
            "term_id" : "HP:0003581"
        },
        "age" : "P17Y0M"
    },
    "bio_characteristics" : [
        {
            "ontology_terms" : [
                {
                    "term_label" : "Glioma",
                    "term_id" : "NCIT:C3059"
                },
                {
                    "term_label" : "Brain NOS",
                    "term_id" : "PGX:ICDOT:C719"
                }
            ],
            "description" : “Juvenile high grade glioma",
        }
    ],
    "external_identifiers" : [
        {
            "database" : "Pubmed",
            "identifier" : "25752754",
            "relation" : "reported_in"
        }
    ],
    "attributes" : {
        "grade" : { "values" : [ { "string_value" : "4" } ] },
        "histone" : { "values" : [ { "string_value" : "wt" } ] },
    }
}
```

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
(LINDA GROB)
SAUMYA GUPTA
(ROMAN HILLJE)
(NITIN KUMAR)
(ALESSIO MILANESE)

SIB

HEINZ STOCKINGER
SÉVERINE DUVAUD
VASSILIOS IOANNIDIS
DANIEL TEIXEIRA

THOMAS EGGERMANN
ROSA NOGUERA
REINER SIEBERT
CAIUS SOLOVAN

GA4GH DWG + CWG

JACQUI BECKMANN
ANTHONY BROOKES
MELANIE COURTOT
MARK DIEKHANS
MELISSA HAENDEL
DAVID HAUSSLER
SARAH HUNT
STEPHEN KEENAN
SUZY LEWIS
DAVID LLOYD
MICHAEL MILLER
HELEN PARKINSON
GUNNAR RÄTSCH
ELEANOR STANLEY
DAVID STEINBERG
JULIA WILSON

ELIXIR & CRG

JORDI RAMBLA DE ARGILA
S. DE LA TORRE PERNAS
SUSANNA REPO
SERENA SCOLLEN

University of Zurich[UZH]

SIB

elixir

Global Alliance for Genomics & Health