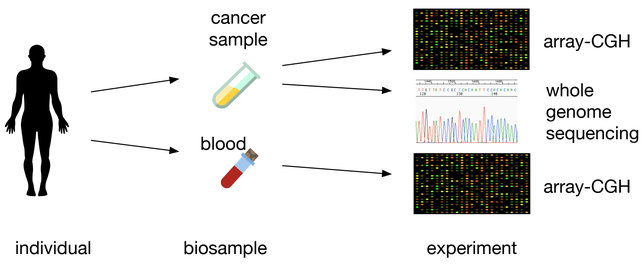
Biocuration to Cancer Genomics: Resources, Standards, Data Science

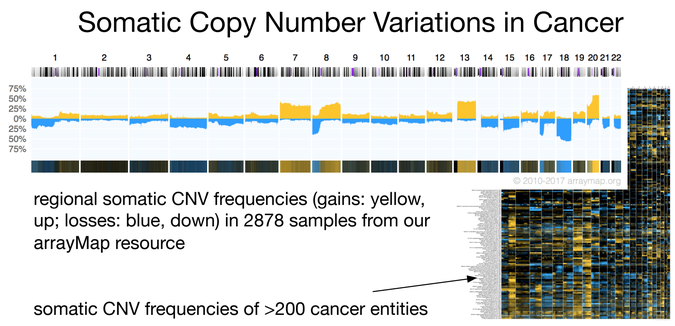
Theoretical Oncogenomics Group - Michael Baudis, University of Zurich & SIB

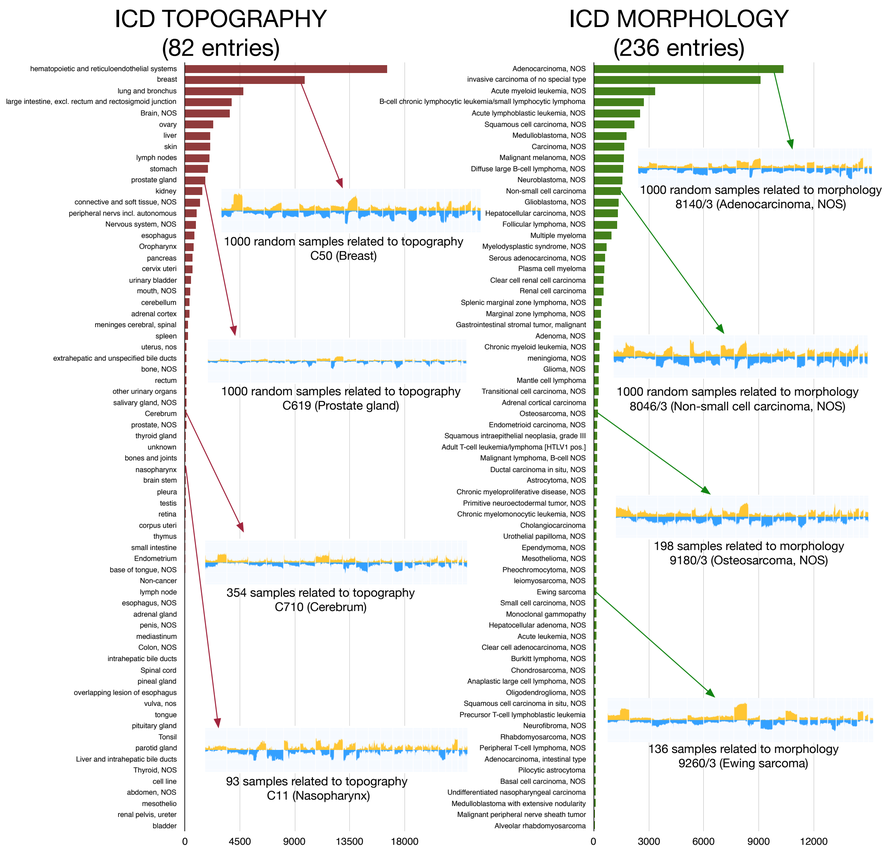
Curated Cancer Genome Data Resources: arrayMap & Progenetix

Screening for somatic mutations in cancer is integral to diagnostic and target evaluation for personalized therapeutic approaches. arrayMap is a curated oncogenomic resource, focusing on copy number aberration (CNA) profiles derived from genomic arrays based on raw data from NCBI's Gene Expression Omnibus (GEO), EBI's ArrayExpress and through mining of publication data. Whereas arrayMap represents data down to probe-level annotations, the parental *Progenetix* resource provides annotated genome variant analysis from additional sources and serves as metadata reference.



Object hierarchy of the GA4GH data model in arraymap and progenetix





Our group is involved in developing and implementing the hierarchical schema of the Global Alliance for Genomics and Health (GA4GH). Its representation of the data for individuals, biosamples and experiments allows for further meta-analysis on different knowledge levels.

For more than 60'000 cancer related samples, we have manually assessed the anatomical site of origin (ICD TOPOGRAPHY) and the characteristic of the tumor histology (ICD MORPHOLOGY). An analysis of the distribution of diagnostic classes highlights strong biases in cancer sudies. The data shows a representation of 26% of the total cancer types and subtypes proposed in ICD-O-3, where half of the data focuses on 7% of the sites.

Under an epistemologic paradigm our data collections reflect knowledge gaps in the cancer genome research landscape, and highlight geographic biases which will be able to guide the direction of future studies.

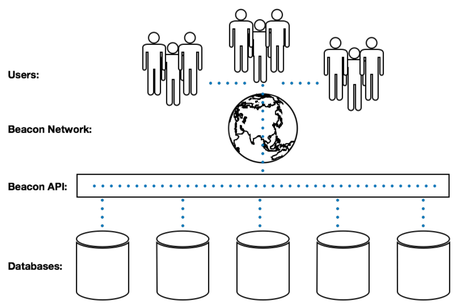
Beacon+: Genomic Data Discovery Based on GA4GH Data Models

Overview

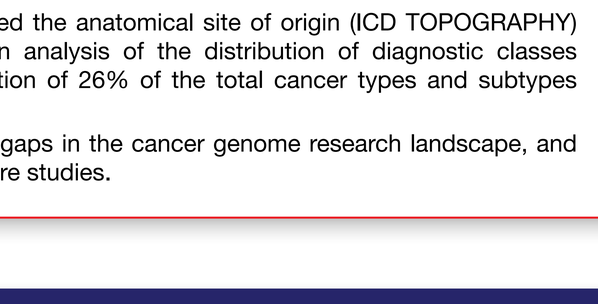
The Beacon Project of the Global Alliance for Genomics and Health (GA4GH) tests the willingness of international data portals to share human genomic information, and to solve technical and regulatory challenges in their local environments. The original Beacon protocol was extremely light-weight, simply allowing queries for existence of a given SNV, in a static dataset. So far, over 50 genome data centers from all over the world have linked themselves into the Beacon Network. In an ELIXIR project to advance Beacon development, our group at the University of Zurich is driving the extension of the Beacon protocol.

Beacon+

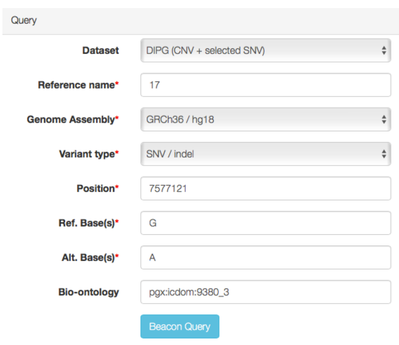
- Extending beacon queries beyond SNP calls and yes/no response
- Two new queries types: structal variants and metadata (i.e. ontology terms for disease selection)
- Quantitative responses (e.g. variant frequency per diagnostic entity)
- Direct implementation of a GA4GH compatible data model on the server, as demonstrator of the schema feasibility in a production setting
- Cancer Beacon+ prototype backed by arrayMap (>5Mill. structural variants)
- Original research data sets with different types of variants (i.e. "DIPG" childhood gliomas; MacKay *et al*., Cancer Cell 2017)
- Developing a query paradigm for range queries and structural genome variants, using a "fuzzy" matching approach



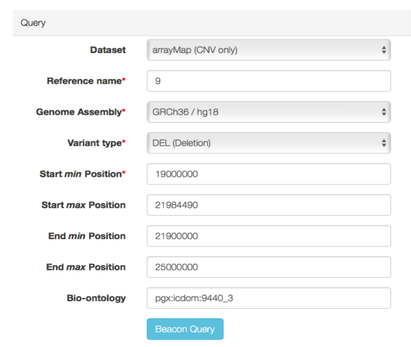
The Beacon Network



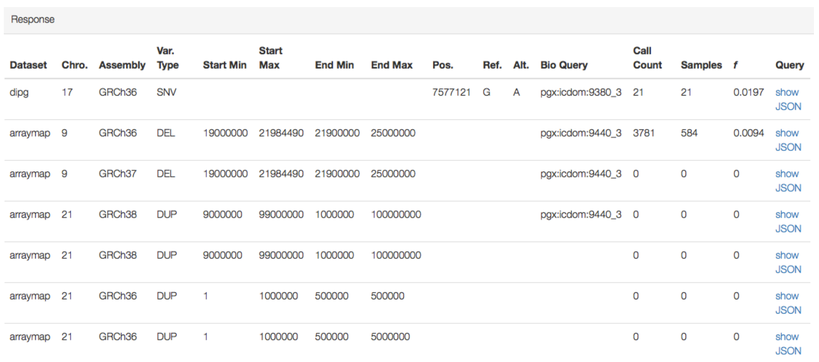
SNV Query



Strucutal Variant Query



Search resaults



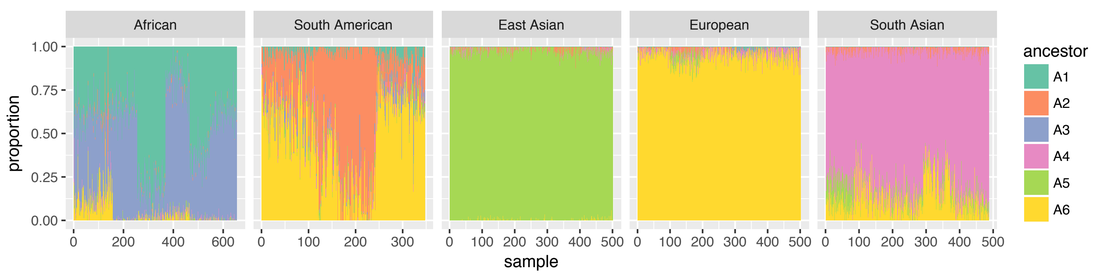
Response

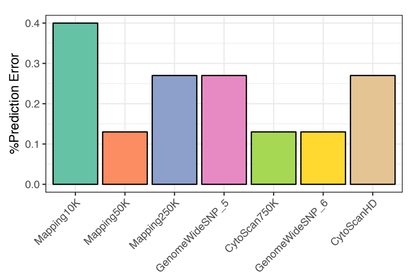
Dataset	Chro.	Assembly	Var. Type	Start Min	Start Max	End Min	End Max	Pos.	Ref.	Alt.	Bio Query	Call Count	Samples	f	Query
dipg	17	GRCh36	SNV					7577121	G	A	pgx:icdom:9380_3	21	21	0.0197	show JSON
arraymap	9	GRCh36	DEL	19000000	21984490	21900000	25000000				pgx:icdom:9440_3	3781	584	0.0094	show JSON
arraymap	9	GRCh37	DEL	19000000	21984490	21900000	25000000				pgx:icdom:9440_3	0	0	0	show JSON
arraymap	21	GRCh38	DUP	9000000	99000000	1000000	100000000				pgx:icdom:9440_3	0	0	0	show JSON
arraymap	21	GRCh38	DUP	9000000	99000000	1000000	100000000					0	0	0	show JSON
arraymap	21	GRCh36	DUP	1	1000000	500000	500000					0	0	0	show JSON
arraymap	21	GRCh36	DUP	1	1000000	500000	5000000					0	0	0	show JSON

Studying population-specific molecular patterns in cancer genomes

Malignant neoplasias are based on the accumulation of mutations in cells during the lifetime of an individual ("somatic mutations"), which can be influenced by inherited ("germline") genome variations. As tumor types and incidences differ among human populations, the genetic background of individuals could be one factor influencing somatic variation and subsequent tumorigenesis. Most cancer genome studies have been conducted on individual tumor types and cohorts of similar genomic backgrounds, and the systematic analyses and integration of multiple available data sources are lacking.

Here, we perform a meta-analysis of the curated oncogenomic data from the arrayMap database, derived from various types of genomic arrays, and combine genomic profiles with epidemiological data to evaluate the population specificity of genome variations in cancer.





From sequencing data of 2504 individuals over 5 super-populations from 1000Genome project, we extract the SNP markers corresponding to Affymetrix platforms and use them for subsequent sample analysis. First, we show that using admixture analysis, the population classification is accurate even from low-resolution arrays (10k markers). This will append genome-derived population information to the Progenetix database, as an addition layer to the geographic location of the publication-affiliated institute. As next step, we will link different types of chromosomal aberration (e.g. cn-LOH) to the identified population group for over 46,000 cancer samples to discover potential population-specific oncogenic patterns.

Total 58290 samples, incl. 46599 tumor samples