A Reference Resource for Copy Number Variations in Cancer
Implementing GA4GH Standards to Drive an Open Oncogenomics Resource
Department of Molecular Life Sciences

Genome screening at the core of “Personalized Health”

- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications

- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**

- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts

- Our work @ UZH:
  - **cancer** genome repositories
  - biocuration
  - protocols & formats

---

**Genomic Array Technology**

**Geographic distribution of 104,543 genomic array, 36,766 chromosomal CGH and 15,409 whole genome/exome based cancer genome datasets**

---

**Figure:** The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.
Genomic alterations in Cancer

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations (CNA)**
- Epigenetic changes (e.g. DNA methylation abnormalities)
History & Current State...
Origins & Trajectory of the Progenetix Resource
Progenetix, early 2000's

Text conversion for CNVs

- originally an internal, to store CGH profiles at Peter Lichter’s group at the German Cancer Research Center (DKFZ), starting in 1998
- expansion to include literature derived data, with a focus on malignant non-Hodgkin’s lymphomas
- based on articles and supplements with cytoband-based rev ish CGH results
- sometimes rich, but unstructured associated information
- PDFs readable, but not well suited for data extraction (character entities, text flow)

Example input Data: Korn et al., 1999 Genes, chromosomes and cancer
**arrayMap (2012 - 2020)**

**Probe-Level Genomic Array Data in Cancer**

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 oncomicron CNA data.

The current data reflects:
- 72724 genomic array profiles
- 898 experimental series
- 257 array platforms
- 341 ICD-O cancer entities
- 795 publications (PubMed entries)

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 Cytochromes and Oncogenicities" at the Institute of Molecular Life Sciences of the University of Zurich.

**RELATED PUBLICATIONS**


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.

© 2000 - 2019 Michael Baudis, refreshed 2019-06-12721/00/192 in 6.00s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.
Progenetix in 2021

Cross-platform Oncogenomics

- source data (i.e. array probe data access) and annotation derived (aCGH, WGS, WES, other arrays)
- >130'000 cancer and reference CNA profiles
- systematic metadata annotations following GA4GH standards
- unrestricted access w/o registration
- data access API
- online visualization
- CNA statistics

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on individual sample data from currently 139448 samples.

Example for aggregated CNV data in 10324 samples in Malignant Thoracic Neoplasm. Here the frequency of regional copy number gains and losses are displayed for all 22 autosomes.

Local CNV Frequencies

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

Cancer CNV Profiles

The progenetix resource contains data of 788 different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [Publications] page Progenetix provides 4025 annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.
Progenetix Implements & Drives the GA4GH Beacon Protocol

CNV Queries, Ontology Based Filters, Data Delivery through Handovers...
Progenetix & Beacon

Demonstrator for GA4GH Beacon based genomic reference resource

- the CNV content of Progenetix has been a driver to develop the range and bracket variant query options
- extensive sample annotations using CURIEs with hierarchical ontologies for "biocharacteristics" (NCIT ...) and external references (cellosaurus, geo, PMID ...) serve implementation scenarios for Beacon testing and "production" environment
- custom implementation
  - "bycon" code on GH > /progenetix/bycon/
  - MongoDB backend
  - React front-end on GH > /progenetix/progenetix-web/
Beacon v2: Extended Variant Queries
Range and Bracket queries enable positional wildcards and fuzziness

- Genome Range Queries provide a way to "fish" for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, reference or alternate bases) limit the scope of the responses
- new Beacon v2 size parameters to limit structural variants (e.g. "focal" CNVs)

- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g. the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions
Beacon v2 Filters

Example: Use of hierarchical classification systems (here NCIt neoplasm core)

- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications
- implicit OR with otherwise assumed AND
- implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations
- data handover (Beacon v1.1+) enables further data exploration and export scenarios
Progenetix & Beacon v1->2

Handover elements in Beacon responses

- Progenetix utilizes handovers to deliver data matched by the Beacon queries
- These handovers are interpreted by the front end to populate different parts of the UI, w/o the need of active selection
- Handovers are either standard Beacon v2 paths or dedicated custom functions
DATA PIPELINE

BIOCURATION

BIOINFORMATICS

arrayMap

progenetix
Metadata Curation

**Standardized, machine-readable**

- Developed by international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI)
- Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations
- Representing varying granularity

<table>
<thead>
<tr>
<th>Subsets</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI/3262: Neoplasm</td>
<td>88844</td>
</tr>
<tr>
<td>NCI/3263: Neoplasm by Site</td>
<td>84747</td>
</tr>
<tr>
<td>NCI/156482: Genitourinary System Neoplasm</td>
<td>11616</td>
</tr>
<tr>
<td>NCI/156483: Benign Genitourinary System Neoplasm</td>
<td>219</td>
</tr>
<tr>
<td>NCI/4893: Benign Urinary System Neoplasm</td>
<td>90</td>
</tr>
<tr>
<td>NCI/4778: Benign Kidney Neoplasm</td>
<td>90</td>
</tr>
<tr>
<td>NCI/159290: Kidney Leiomyoma</td>
<td>1</td>
</tr>
<tr>
<td>NCI/4526: Kidney Oncocytoma</td>
<td>82</td>
</tr>
<tr>
<td>NCI/8383: Kidney Adenoma</td>
<td>7</td>
</tr>
<tr>
<td>NCI/7677: Benign Reproductive System Neoplasm</td>
<td>129</td>
</tr>
<tr>
<td>NCI/4934: Benign Female Reproductive System Neoplasm</td>
<td>129</td>
</tr>
<tr>
<td>NCI/2895: Benign Ovarian Neoplasm</td>
<td>68</td>
</tr>
<tr>
<td>NCI/4510: Benign Ovarian Epithelial Tumor</td>
<td>68</td>
</tr>
<tr>
<td>NCI/4039: Benign Ovarian Mucinous Tumor</td>
<td>68</td>
</tr>
<tr>
<td>NCI/4512: Ovarian Mucinous Cystadenoma</td>
<td>58</td>
</tr>
<tr>
<td>NCI/4060: Ovarian Cystadenoma</td>
<td>58</td>
</tr>
<tr>
<td>NCI/4512: Ovarian Mucinous Cystadenoma</td>
<td>58</td>
</tr>
<tr>
<td>NCI/3609: Benign Uterine Neoplasm</td>
<td>71</td>
</tr>
<tr>
<td>NCI/3608: Benign Uterine Corpus Neoplasm</td>
<td>71</td>
</tr>
<tr>
<td>NCI/3434: Uterine Corpus Leiomyoma</td>
<td>71</td>
</tr>
<tr>
<td>NCI/156484: Malignant Genitourinary System Neoplasm</td>
<td>11171</td>
</tr>
<tr>
<td>NCI/157774: Metastatic Malignant Genitourinary System Neoplasm</td>
<td>2</td>
</tr>
<tr>
<td>NCI/146893: Metastatic Genitourinary System Carcinoma</td>
<td>2</td>
</tr>
<tr>
<td>NCI/8946: Metastatic Prostate Carcinoma</td>
<td>2</td>
</tr>
<tr>
<td>NCI/164141: Genitourinary System Carcinoma</td>
<td>10561</td>
</tr>
</tbody>
</table>

"Colorectal Cancer" or "Rectal Mucinous Adenoca."
Beacon & Phenopackets

Data discovery and delivery using standardized GA4GH formats and schemas

- modern standards and protocols are essential for federation and exchange of biomedical data
- emerging/established principles are the use of hierarchical coding systems and with widespread use of CURIEs
- Non-categorical annotations e.g. stage, grade, followup time, survival status, genomic sex, age at diagnosis based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)

➡ Beacon query filters correspond well to Phenopackets data

➡ Phenopackets as supported protocol for Beacon data delivery

```json
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},
"material" : {
  "id" : "EFO:0009656",
  "label" : "neoplastic sample"
},
{ "ageAtDiagnosis" : "P25Y3M2D"
},
{ "sampled_tissue" : {
  "id" : "UBERON:0002037",
  "label" : "cerebellum"
}
},
{ "histological_diagnosis" : {
  "id" : "NCIT:C3222",
  "label" : "Medulloblastoma"
}
}
```
Progenetix is a Cancer CNV Reference Resource
Cancer CNV Profiles, Gene CNV Frequencies, Cohorts & Publications
Progenetix in 2021

Cancer type representation

- Hierarchical aggregation
- CNV landscape overview
- Sample retrieval

Cancer Types

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated data is available by clicking the code. Additionally, a selection of the corresponding samples can be initiated by clicking the sample number or selecting one or more classes through the checkboxes.

Sample selection follows a hierarchical system in which samples matching the child terms of a selected class are included in the response.

Cancer Classification: NCIT Cancer Core

Filter subsets ...

Hierarchy Depth: collapsed

- NCIT:C3010: Endocrine Neoplasm (3319 samples)
- NCIT:C3030: Eye Neoplasm (280 samples)
- NCIT:C3052: Digestive System Neoplasm (15194 samples)
- NCIT:C3077: Head and Neck Neoplasm (3769 samples)
- NCIT:C3268: Nervous System Neoplasm (16270 samples)
- NCIT:C2963: Cranial Nerve Neoplasm (19 samples)
- NCIT:C3321: Peripheral Nervous System Neoplasm (901 samples)
- NCIT:C35562: Neuroepithelial, Perineurial, and Schwann Cell Neoplasm (11690 samples)
- NCIT:C3787: Neuroepithelial Neoplasm (11214 samples)
- NCIT:C3059: Glioma (8888 samples)
- NCIT:C3716: Primitive Neuroectodermal Tumor (2213 samples)
- NCIT:C6963: Neuroblastoma (2001 samples)
Glioblastoma

Assembly: GRCh38 Gene: CDKN2A Filters: NCIT:C3058

Samples: 2426  Variants: 1733  Calls: 2566

Table:

<table>
<thead>
<tr>
<th>Matched Subset Codes</th>
<th>Subset Samples</th>
<th>Matched Samples</th>
<th>Subset Match Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>icdot-C71.8</td>
<td>1</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>icdot-C71.4</td>
<td>4</td>
<td>3</td>
<td>0.760</td>
</tr>
<tr>
<td>UBERON:0002021</td>
<td>4</td>
<td>3</td>
<td>0.750</td>
</tr>
<tr>
<td>icdom-94403</td>
<td>4291</td>
<td>2396</td>
<td>0.558</td>
</tr>
<tr>
<td>NCIT:C3058</td>
<td>4375</td>
<td>2396</td>
<td>0.548</td>
</tr>
<tr>
<td>NCIT:C3796</td>
<td>84</td>
<td>30</td>
<td>0.357</td>
</tr>
<tr>
<td>Int. ID</td>
<td>Digest</td>
<td>Biosample</td>
<td>Chr.</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------</td>
<td>--------------------</td>
<td>------</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b9478</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftva5zv</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b9493</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftva601</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b94a4</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftva603</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b94ae</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftva606</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b94b3</td>
<td>9:0-140273252:DEL</td>
<td>pgxbs-kftva608</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576a727983b2e00b94be</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftva60b</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576d727983b2e00bae65</td>
<td>9:0-40199999:DEL</td>
<td>pgxbs-kftvgl7e</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576d727983b2e00bae7f</td>
<td>9:0-32799999:DEL</td>
<td>pgxbs-kftvgl7j</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576d727983b2e00bae87</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftvgl7t</td>
<td>9</td>
</tr>
<tr>
<td>pgxvar-5bab5b576d727983b2e00bae9b</td>
<td>9:0-51799999:DEL</td>
<td>pgxbs-kftvgl7u</td>
<td>9</td>
</tr>
</tbody>
</table>
Progenetix API

Data & Plots

- Through multiple API endpoints
- Segmented CNV data in .pgxseg (tabular) and JSON format
- CNV frequency histograms for disease, study or cohort
- R package for circos, heatmap and more...

```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```
Progenetix
Services, Documentation...

- services e.g. for disease code translation (NCIt <=> ICD-O; UBERON ...)
- documentation of data schema, API ...

Welcome to the Progenetix documentation pages

The Progenetix Resource Documentation provides information and links related to the Progenetix cancer genome resource and the related Progenetix code repositories contains projects, such as data conversion scripts, ontology mappings and code for the Beacon project.

Progenetix Website Code Repositories
- Progenetix Source Code
- Related Projects

Latest News

Progenetix File Formats

Standard Progenetix Segment Files: pgxseg

Progenetix uses a variation of a standard tab-separated column text file such as produced by array or sequencing CNV software, with an optional metadata header for e.g. pilot or grouping instructions.

@mbauds 2021-02-22: more ...

Beacon+ and Progenetix Queries by Gene Symbol

We have introduced a simple option to search directly by Gene Symbol, which will match to any genomic variant with partial overlap to the specified gene. This works by expanding the Gene Symbol (e.g. TP53, CDKN2A ...) into a range query for its genomic coordinates (maximum 200). Such queries - which would return all whole-chromosome CNV events covering the gene of interest - should be narrowed by providing e.g. Variant Type and Maximum Size (e.g. 2000000) values.

@mbauds 2021-02-22: more ...

The Progenetix oncogenic resource in 2021

Qingyan Huang, Paula Carrio Corbo, Wei Gao, Rahel Palaos, Michael Baudis

Link: https://doi.org/10.1101/2021.02.18.420227

This article provides an overview of recent changes and additions to the Progenetix database and the services provided through the resource.

2021-02-15: more ...

Diffuse Intrinsic Pontine Glioma (DIPG) cohort

Diffuse Intrinsic Pontine Glioma (DIPG) is a highly aggressive tumor type that originate from glial cells in the pons area of brainstem, which controls vital functions including breathing, blood pressure and heart rate. DIPG occurs frequently in early childhood and has a 5-year survival rate below 1 percent. Progenetix has now incorporated the DIPG cohort, consisting of 1007 individuals from 16 publications. The measured data include copy number variation as well as in (p)art) point mutations on relevant genes, e.g. TP53, NF1, ARTR, TRA7, PRAG. promoter.

@mbauds 2021-02-15: more ...

arrayMap is Back

After some months of dormancy, the arrayMap resource has been relaunched through integration with the new Progenetix site. All of the original arrayMap data has now been integrated into Progenetix, and of today the arrayMap AngioMap domain maps to a standard Progenetix search page, where only data samples with existing source data (e.g. probes specific array files) will be presented.

@mbauds 2021-02-06: more ...
Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".
Beacon API Leads

- Jordi Rambla
- Anthony Brooks
- Juha Törnroos
- Discovery WS
  - Michael Baudis (Beacon)
  - Marc Fiume (Networks)
- ELIXIR
  - Gary Saunders
  - David Lloyd
  - Serena Scollen

Sabela de la Torre Pernas

Beaudisgroup @ UZH
(Ni Ai)
- Michael Baudis (Haoyang Cai)
  - (Paula Carrio Cordo)
- Bo Gao
- Qingyao Huang (Saumya Gupta)
  - (Nitin Kumar)
- Sofia Pfund
- Rahel Paloots
- Ziying Yang
- Hangjia Zhao
- Pierre-Henri Toussaint

{S}{B} and GA4GH
- Melanie Courtot
- Helen Parkinson
  - many more ...

beacon-project.io

beacon.progenetix.org/ui/

github.com/ga4gh-beacon/