Progenetix & GA4GH
A cancer genomics resource built around and driving GA4GH standards
Genome screening at the core of “Personalised 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
Types of genomic alterations in Cancer

Imbalanced Chromosomal Changes: CNV

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

[Diagram showing various chromosomal alterations, including Polyploidy, Aneuploidy, Uniparental disomy, Non-reciprocal translocation, Partial deletion, Partial duplication, Amplification (double minutes), and Amplification (HSR).]

Grade et al., 2015 Recent Results Cancer Res
Gain of chromosome arm 13q in colorectal carcinoma

MYCN amplification in neuroblastoma (GSM314026, SJN88_N cell line)

low level/high level copy number alterations (CNAs)

2-event, homozygous deletion in a Glioblastoma
Quantifying Somatic Mutations In Cancer

On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

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))

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

GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER
History & Current State...

Origins & trajectory of the Progenetix Resource
A compilation of published CGH analyses with reported case and band specific results

Currently includes 2613 cases from 92 publications

Automatic conversion of ISCN format to aberration matrix with 393 bands resolution

Online CGH database

2613 neoplasias

787 lymphoid

451 NHL

Material and Methods

Material and Methods

Chromosomal aberration data of more than 5478 cases from 196 publications describing results of Comparative Genomic Hybridisation (CGH) experiments were collected. Minimal requirements were disposition of a malignant or benign neoplasia, analysis of clinical tumor samples and report of the analysis results on a case by case basis, reviewed to the level of single chromosomal bands. Data was transformed from the diverse annotation of the original data into the common format of the aberration matrix for the next generation of data mining tools. The raw data was stored in the form of 2-dimensional matrices containing data about the gains or losses of specific bands (14). For the analysis of the non-linear ISCN data in a 4-dimensional matrix, a reverse pattern matching algorithm was developed in Part. Graphical representations and cluster schemes are generated for all different subsets (Pathologies, ICD-O-3 entities, meta-groups) and presented on the progenetix.net website.

Results

Out of 4896 tumor samples, 5933 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.0% for a 95% score, 12.9% at 90% and 56.0% for a gain (max. 15.6% at 8p23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (19 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (20), neuroendocrine tumors (23.3% with max. 86.2% at 3p14) and phaeochromocytomas (70.0%, max. 99.9% at 9q34). Prominent gain regions were common to all subsets with 31.0% frequency, followed by lymphoid tumors (32.9%, max. 95.5%, 11q13), T-PLL, (54.3%, max. 81.0% at 14q32) and diffuse large B-cell lymphomas (28.2%, max. 95.2%, 8q24). Malignant tumors with more than 100 cases were lymphoid (31.6%), melanomas (22.3%), breast cancer (21.3%), sarcomas (19.5%), renal cell carcinoma (16.7%), adenocarcinomas (15.3%) and prostate (12.2%).

Examples of tumor genotypes

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Gains</th>
<th>Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLL</td>
<td>+12</td>
<td></td>
</tr>
<tr>
<td>Extraneous NHL</td>
<td>+12</td>
<td></td>
</tr>
<tr>
<td>MBL</td>
<td>+3q</td>
<td>+8</td>
</tr>
<tr>
<td>CLL/Extraneous NHL</td>
<td>+8</td>
<td>+12</td>
</tr>
<tr>
<td>MZBCL</td>
<td>+9p</td>
<td>-8</td>
</tr>
<tr>
<td>MCL</td>
<td>-8</td>
<td>+3q</td>
</tr>
<tr>
<td>T-PLL</td>
<td>+3q</td>
<td>+8</td>
</tr>
<tr>
<td>NHL</td>
<td>+3q</td>
<td>-8</td>
</tr>
</tbody>
</table>

Conclusions

The progenetix.net project was able to:
1. collect a large dataset of genomic aberration data generated through a molecular cytogenetic screening techniques (CGH).
2. develop software tools to transform these data into a meta format compatible to community used generic interface descriptions.
3. produce graphical and numerical output from these data for better extraction and statistical analysis.

For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and to further analyze these data and find new clinical relevant pathways in the oncogenesis of different tumor entities. The transformed raw data of the progenetix.net collection is available for research purposes over the website.

Michael Baudis | Presentation ASH | Orlando 2001-12-05

Michael Baudis | BCATS Biocomputing at Stanford II | Stanford Nov 2002
Progenetix Database in 2003

Text conversion for CNVs

- 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)

<table>
<thead>
<tr>
<th>Case</th>
<th>Gain in common</th>
<th>Loss in common</th>
<th>Primary tumor only</th>
<th>Metastasis only</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>7, 8q24-qter, 13q11-qter, 20q11-qter, Xq11-Xter</td>
<td>1p32-qter, 2p21-qter, 4q24-qter, 15q11-q15, 17p11-qter, 18p11-qter</td>
<td>11p11-pter</td>
<td>12+</td>
</tr>
<tr>
<td>145</td>
<td>4q28-q8, 6p11-p13, 8p11-q12, 9p21-qter</td>
<td>1q11-qter, 4q31-qter, 6q11-qter, 8p12-p1ter, 11p11-qter, 16q11-qter, 17p11-p1ter, 18p11-qter</td>
<td>21q11-qter</td>
<td>13q21-p1ter, 20p11-p1ter</td>
</tr>
<tr>
<td>53</td>
<td>7, 8q11-qter, 8q23-qter, 13q11-qter, 20p11-p12, 20q11-qter</td>
<td>14p14-qter, 4p21-qter, 8p12-p1ter, 16q14-qter, 18q11-qter, 20p11-qter</td>
<td>8p11-qter, 11p13-q14, 14q11-qter</td>
<td>11q22-qter, 16p11-qter, 17q11-qter, 19p11-qter, 21q11-qter, 22q11-qter</td>
</tr>
<tr>
<td>147</td>
<td>7, 13q11-qter, 20q11-qter</td>
<td>8p11-qter, 18p14-qter, 4q28-qter, 8p11-qter, 17q11-q2, 21q11-qter</td>
<td>11p11-pter, 16q11-qter, 19p11-qter, 21q11-qter, 22q11-qter</td>
<td>11q22-qter, 16p11-qter, 19p11-qter, 21q11-qter, 22q11-qter</td>
</tr>
</tbody>
</table>

**Table 1. Clinical Data**

<table>
<thead>
<tr>
<th>Case number</th>
<th>Age</th>
<th>Sex</th>
<th>Site</th>
<th>Stage Grades *</th>
<th>Diagnosis of metastatic disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>40</td>
<td>M</td>
<td>Transverse colon</td>
<td>IV 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>6</td>
<td>79</td>
<td>M</td>
<td>Ascending colon</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>11</td>
<td>56</td>
<td>M</td>
<td>Transverse colon</td>
<td>II 2</td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>F</td>
<td>Rectosigmoid</td>
<td>IV 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>13</td>
<td>65</td>
<td>M</td>
<td>Descending colon</td>
<td>II 9</td>
<td>Synchronous</td>
</tr>
<tr>
<td>14</td>
<td>60</td>
<td>M</td>
<td>Rectum</td>
<td>III 3</td>
<td>Metastatic</td>
</tr>
<tr>
<td>15</td>
<td>51</td>
<td>F</td>
<td>Rectum</td>
<td>III 3</td>
<td>Metastatic</td>
</tr>
<tr>
<td>19</td>
<td>63</td>
<td>M</td>
<td>Rectosigmoid</td>
<td>III 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>20</td>
<td>63</td>
<td>M</td>
<td>Rectum</td>
<td>III 9</td>
<td>Metastatic</td>
</tr>
<tr>
<td>21</td>
<td>64</td>
<td>F</td>
<td>Sigmoid colon</td>
<td>IV 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>22</td>
<td>71</td>
<td>M</td>
<td>Rectum</td>
<td>III 3</td>
<td>Metastatic</td>
</tr>
<tr>
<td>49</td>
<td>72</td>
<td>M</td>
<td>Cecum</td>
<td>IV 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>53</td>
<td>72</td>
<td>F</td>
<td>Sigmoid colon</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>104</td>
<td>61</td>
<td>M</td>
<td>Sigmoid colon</td>
<td>IV 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>105</td>
<td>58</td>
<td>M</td>
<td>Ascending colon</td>
<td>II 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>107</td>
<td>77</td>
<td>F</td>
<td>Cecum</td>
<td>IV 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>108</td>
<td>53</td>
<td>F</td>
<td>Sigmoid flexure</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>112</td>
<td>68</td>
<td>M</td>
<td>Rectum</td>
<td>III 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>113</td>
<td>41</td>
<td>M</td>
<td>Sigmoid flexure</td>
<td>IV 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>114</td>
<td>49</td>
<td>M</td>
<td>Sigmoid flexure</td>
<td>IV 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>116</td>
<td>73</td>
<td>M</td>
<td>Rectosigmoid</td>
<td>III 3</td>
<td>Synchronous</td>
</tr>
<tr>
<td>120</td>
<td>24</td>
<td>F</td>
<td>Descending colon</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>123</td>
<td>62</td>
<td>F</td>
<td>Rectum</td>
<td>III 2</td>
<td>Metastatic</td>
</tr>
<tr>
<td>124</td>
<td>42</td>
<td>M</td>
<td>Rectum</td>
<td>IV 9</td>
<td>Synchronous</td>
</tr>
<tr>
<td>145</td>
<td>70</td>
<td>M</td>
<td>Rectosigmoid</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
<tr>
<td>147</td>
<td>86</td>
<td>F</td>
<td>Cecum</td>
<td>IV 2</td>
<td>Synchronous</td>
</tr>
</tbody>
</table>

*NCCN/ASCO staging system (Hutter and Sollin, 1996).
*Grade of primary tumor: 1-3, low; 4-6, intermediate; 7-9, high; grade 0, grading unknown.
*Synch = synchronous; diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metasynch = metastatic, diagnosis of metastatic disease after 12 months or later.
Progenetix: Data Scopes
Biomedical and procedural "Meta"data types

• Diagnostic classification
  • mapping text-based cancer diagnoses to standard classification systems

• Provenance data
  • store identifier-based pointers
  • geographic attribution (individual, biosample, experiment)

• Clinical information
  • core set of typical cancer study values:
    → stage, grade, followup time, survival status, genomic sex, age at diagnosis
  • balance between annotation effort and expected usability
Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```bash
$SAMPLE = GSM174832
$Sample_title = 9194
$Sample_geo_accession = GSM174832
$Sample_status = Public on May 01 2007
$Sample_submission_date = Mar 13 2007
$Sample_last_update_date = Mar 13 2007
$Sample_type = genomic
$Sample_channel_count = 1
$Sample_source_name_ch1 = Bone marrow with 96% blasts
$Sample_organism_ch1 = Homo sapiens
$Sample_taxid_ch1 = 9606
$Sample_characteristics_ch1 = common ALL; Age: 9.2 yrs; Gender: F
$Sample_molecule_ch1 = genomic DNA
$Sample_extract_protocol_ch1 = QiaAmp purification kit (Qiagen)
$Sample_label_protocol_ch1 = Biotinylated DNA was prepared according to the standard Affymetrix protocol for 100k assay (Rev. 3.1) or 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3, Affymetrix).
$Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol for 100k assay (Rev. 3.1) or 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3, Affymetrix) using an Affymetrix hybridisation oven 640 and an Affymetrix Fluidic station 450.
$Sample_scan_protocol = Scanning performed according to the standard Affymetrix protocol for 100k assay (Rev. 3.1) or 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3, Affymetrix) using an Affymetrix scanner 3000.
$Sample_description = primary ALL diagnosis sample
$Sample_data_processing = copy number detection using CNAG2.0 software (http://www.genome.umin.jp/)
$Sample_platform_id = GPL3718
$Sample_contact_name = Roland,P., Kuiper
$Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
$Sample_contact_phone = +31243610868
$Sample_contact_fax = +31243668752
$Sample_contact_department = Human Genetics
$Sample_contact_institute = Radboud University Nijmegen Medical Centre
$Sample_contact_address = Geert Grooteplein 10
$Sample_contact_zip/postal_code = 6525GA
$Sample_contact_country = Netherlands
$Sample_series_id = GSE7255
```

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

SAMPLE = GSM286922
Sample_title = 481 - mAdID:75320
Sample_geo_accession = GSM286922
Sample_status = Public on Sep 04 2008
Sample_submission_date = May 06 2008
Sample_last_update_date = Nov 26 2008
Sample_type = genomic
Sample_channel_count = 2
Sample_source_name_ch1 = Normal Lymphocytes
Sample_organism_ch1 = Homo sapiens
Sample_taxid_ch1 = 9606
Sample_characteristics_ch1 = Tissue: lymphocytes
Sample_characteristics_ch1 = Disease state: Lymphoma
Sample_characteristics_ch1 = Individual: 481
Clinical info: Submitting diagnosis: DLBCL
Clinical info: Follow up status: ALIVE
Clinical info: Follow up years: 10.75
Clinical info: Chemotherapy: CHOP-Like Regimen
Clinical info: LDH ratio: 0.82
Clinical info: Number of extranodal sites: 1

Gender: male
Tissue: lymph node
Disease state: Lymphoma
Individual: 481
Clinical info: Submitting diagnosis: DLBCL
Clinical info: Follow up status: ALIVE
Clinical info: Follow up years: 10.75
Clinical info: Chemotherapy: CHOP-Like Regimen
Clinical info: LDH ratio: 0.82
Clinical info: Number of extranodal sites: 1

$sample->( samplekey ) = 'DEATH';
$sample->( matches ) = [ { 'death', 'dead', 'vital_status', 'dead_alive', 'alive_deads', } ];
if ( $sample->( retv ) == /[^{]+}$/ ) {
   $sample->( $samplekey ) = $sample->( samplekey )
   $sample->{ $samplekey } = _normDeath($sample->( retv )

Channel 1 is normal -> Cave value swap!
Gender or "chromosomal sex"?

context indicates years, but if it would be a medulloblastoma...
This survival status annotation not known to parser...
The Progenetix oncogenomic resource in 2021

Qingyao Huang1,2, Paula Carrio-Cordo1,2, Bo Gao1,2, Rahel Paloots1,2 and Michael Baudis1,2,*

1Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
2Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland
*Corresponding author: Tel: +41 44 639 34 40; Email: michael.baudis@imls.uzh.ch


Abstract

In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identity and characterization molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 130,683 (including 115,357 tumor) copy number variation (CNA) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNA landscape summary over 51 distinctive National Cancer Institute Therasaurus cancer terms as well as updates in database schemes, and data access including new web front-end and programmatic data access.

Database URL: progenetix.org

Table 1. Statistics of samples from various data resources

<table>
<thead>
<tr>
<th>Data source</th>
<th>GEO</th>
<th>ArrayExpress</th>
<th>cBioPortal</th>
<th>TCGA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of studies</td>
<td>898</td>
<td>51</td>
<td>38</td>
<td>33</td>
<td>1939</td>
</tr>
<tr>
<td>No. of samples</td>
<td>63,568</td>
<td>4351</td>
<td>19,712</td>
<td>22,142</td>
<td>138,663</td>
</tr>
<tr>
<td>Tumor</td>
<td>52,090</td>
<td>3897</td>
<td>19,712</td>
<td>11,092</td>
<td>115,357</td>
</tr>
<tr>
<td>Normal</td>
<td>11,478</td>
<td>464</td>
<td>0</td>
<td>11,052</td>
<td>23,306</td>
</tr>
</tbody>
</table>

Classifications

| ICD-O (Topography) | 100 | 34 | 88 | 157 | 209 |
| ICD-O (Morphology) | 246 | 908 | 265 | 140 | 491 |
| NCIt | 346 | 148 | 422 | 182 | 788 |

Collections

| Individuals | 63,568 | 4351 | 19,712 | 10,995 | 127,549 |
| Biosamples | 63,568 | 4351 | 19,712 | 22,142 | 138,663 |
| Callsetsa | 63,568 | 4351 | 19,712 | 22,142 | 138,930 |
| Variants | 5,514,126 | 118,4170 | 1,778,096 | 2,654,065 | 10,716,093 |

*set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Therasaurus.

Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene’s coding region with ≤ 6 Mb in size. A: Filter for dataset; B: Filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellsource; D: additional filter for geographic location; E: external link to UCSC Genome Browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched biosamples with description, statistics and reference. More detailed biosample information can be viewed through ‘info’ link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.
Progenetix in 2021
Cancer Genomics Reference Resource

- >116'000 cancer CNV profiles, mapped to >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus where appropriate
- core biosample and technical metadata annotations where accessible
- publication database and code mapping services

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.

Breast Cancer by AJCC v6 Stage (NCIt:C30513)

Example for aggregated CNV data in 362 samples in Breast Cancer by AJCC v6 Stage.
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 810 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
Cancer Type CNA Data

- hierarchical aggregation of cancer samples
- pre-computed CNA frequencies for fast overview
- sample retrieval for custom grouping, visualization
Variant and Metadata for Sample Discovery

• positional queries for genomic variants using the GA4GH Beacon protocol

• metadata queries (diagnoses, identifiers, clinical classes ...) using Beacon "filters"

Progenetix in 2021

- Cancer CNV Profiles
- Search Samples
- Studies & Cohorts
- Publication DB
- Services
- Upload & Plot
- Download Data
- Beacon
- Progenetix Info

**Genome Bracket Query**
(full match)
Progenetix in 2021
Query Results and Variant Frequencies

- genomic variant + metadata queries provide relative result counts / frequencies for mapped entities (NCIt, ICD-O ...)
  - disease-specific CNA event scores
- representation of genome-wide CNA frequency profiles / context
- link-outs to download options, subset visualization, sample exploration ...
## Progenetix in 2021

### Query Results and Variant Frequencies

<table>
<thead>
<tr>
<th>Biosample Id</th>
<th>Dx Classifications</th>
<th>Identifiers</th>
<th>CNV Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>pgbds-kftvgdk8h</td>
<td>NCI:T:C058</td>
<td>PMID:23079654 Sturr D, Witt H et al. (2012): Hotspot mutations in H3F3A and IDH1...</td>
<td>0.26</td>
</tr>
<tr>
<td>icdot-C71.0 Cerebrum</td>
<td>Glioblastoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBERON:001869 Cerebral hemisphere</td>
<td>Glioblastoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pgbds-kftvgk90</td>
<td>NCI:T:C058</td>
<td>PMID:23079654 Sturr D, Witt H et al. (2012): Hotspot mutations in H3F3A and IDH1...</td>
<td>0.304</td>
</tr>
<tr>
<td>icdot-C71.0 Cerebrum</td>
<td>Glioblastoma</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Matched Subset Codes**

<table>
<thead>
<tr>
<th>Subset Samples</th>
<th>Matched Samples</th>
<th>Subset Match Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBERON:0002021</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>icdot-C71.4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>icdot-94403</td>
<td>4291</td>
<td>674</td>
</tr>
<tr>
<td>NCI:T:C3058</td>
<td>4375</td>
<td>674</td>
</tr>
<tr>
<td>UBERON:0018525</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>icdot-C71.1</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>UBERON:000955</td>
<td>7284</td>
<td>661</td>
</tr>
<tr>
<td>icdot-C71.9</td>
<td>7282</td>
<td>661</td>
</tr>
<tr>
<td>icdot-94423</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td>NCI:T:C3796</td>
<td>84</td>
<td>4</td>
</tr>
<tr>
<td>UBERON:001869</td>
<td></td>
<td>1712</td>
</tr>
<tr>
<td>icdot-C71.0</td>
<td></td>
<td>1712</td>
</tr>
</tbody>
</table>
Progenetix in 2021
Query Result Visualizations
Progenetix API

Data & Plots

• "all" of the data can be accessed using API calls

• segmented CNV data in .pgxsg (columnar) and JSON format

• histograms for disease, study or cohort from precomputed frequencies - live generated as SVG for embedding with plot options

https://progenetix.org/cgi/PGX/cgi/collationPlots.cgi?datasetIds=progenetix&id=pgxcohort-TCGAcancers&size_plotimage_w_px=800&size_plotarea_h_px=300&value_plot_y_max=60
Progenetix Services, Documentation...

• services e.g. for disease code translation (NCIt <=> ICD-O; UBERON ...)

• API & documentation "progressing" ...
Map of the geographic distribution (by first author affiliation) of the 104,543 genomic array, 36,766 chromosomal CGH and 15,409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3,240 publications registered in the Progenetix database.
Progenetix and GA4GH Beacon
Implementation driven development of a GA4GH standard
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.
A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections.

YES | NO | \0
ELIXIR Beacon Project

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

v1.1 and roadmap

- **structural variations** (DUP, DEL) in addition to SNV
- ... more structural queries (translocations/fusions...)
- Beacon queries as entry for **data handover** (outside Beacon protocol)
- layered authentication system using ELIXIR AAI

ELIXIR Beacon
https://www.elixir-europe.org/about/implementation-studies/beacons
Europe
Champions: Jordi Rambla, Juha Tornroos, Gary Saunders

v2 filters for phenotypic & technical metadata

v2 Extended quantitative responses
- Ubiquitous **deployment** (e.g. throughout ELIXIR network)
Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

• Since 2016 the Progenetix resource has been used to model options for Beacon development
  ‣ 138334 individual samples from 698 cancer types

• The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries

• Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  ‣ downloads
  ‣ visualization
  ‣ use of external services (UCSC browser display...)

progenetix
Beacon v2 Requests

POSTing Queries

- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents
- Final syntax for core parameters still in testing stages

```
{
    "$schema": "beaconRequestBody.json",
    "meta": {
        "apiVersion": "2.0",
        "requestedSchemas": [
            {
                "entityType": "individual",
                "schema": "https://progenetix.org/services/schemas/Phenopacket/"
            }
        ]
    },
    "query": {
        "requestParameters": {
            "datasets": {
                "datasetIds": ["progenetix"]
            }
        },
        "filterLogic": "OR"
    },
    "pagination": {
        "skip": 0,
        "limit": 10
    },
    "filters": [
        { "id": "NCIT:C4536" },
        { "id": "NCIT:C95597" },
        { "id": "NCIT:C7712" }
    ]
}
```
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
Standardized Data

Data re-use depends on standardized, machine-readable metadata

- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards

- emerging / established principles are the use of hierarchical coding systems where individual codes are represented as CURIEs

- other formats for non-categorical annotations based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)

- these standards become pervasive throughout GA4GH's ecosystem (e.g. Phenopackets ...)

```json
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    }
  },
  "ISO-3166-alpha3" : "CHE"
},

{ "age": "P25Y3M2D" }
Standardized formats and data schemas for developing an "Internet of Genomics"

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes launched in December 2018
- documentation and implementation examples provided by GA4GH members
- not a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
  - Beacon as contributor and user
  - 2021: going forward through integration with GA4GH TASC efforts, towards "standards library"
Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths

- Beacon v2 paths are used in the Beacon specification to scope query and delivery
- Progenetix uses a default `/biosamples/ + query` path for its front end queries, and then collection specific methods for data retrieval (see next)
- current implementation addresses a core subset of all options, and evaluates some still moving targets
  - `variants_interpretations`
  - `variant instances versus prototypes`
  - ...
Beacon & Handover

Beacons v1.1 supports data delivery services

Example data backend implemented by Beacon+
beacon.progenetix.org

Beacon Query
allele_request
biosample_request
individual_request
filters

Beacon Response
beacon_response
handover_id

Handover Action
phenopackets
VCF
graphics

Beacon I/O
Handover
Authentication

Global Alliance
for Genomics & Health
SIB
University of Zurich
UZH
elixir
Onboarding
Demonstrating Compliance

• onboarding server run by CRG (EGA team)
• registering the URI of a server's map document will initiate traversal and testing of services
• blueprint for Beacon service registries
• to be used as demonstrator in GA4GH approval process for the Spring 2022 session
beacon-framework-v2

Beacon Framework version 2

Introduction
The GA4GH Beacon specification is composed by two parts:

- the Beacon Framework (in this repo)
- the Beacon Model (in the Models repo)

The Beacon Framework is the part that describes the overall structure of the API.

Bycon - a Python-based environment for the Beacon v2 genomics API

Bycon, at least at its current stage - is a mix of Progenetix (i.e., GA4GH object model derived, MongoDB implemented) - data management, and the implementation of middleware & server for the Beacon API.

More information about the current status of the package can be found in the inline documentation which is also presented in an accessible format on the Progenetix website.
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".
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".
Progenetix Data Use Cases
Genome CNV coverage in Cancer Classes

- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)

- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)

- range of medians 0.001 (CML) - 0.358 (malignant melanomas)
Drivers? Passengers? Markers?
Disentangling CNA Patterns

Ductal Breast Carcinoma

Glioblastoma
Somatic CNVs In Cancer

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 frequent 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.
Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles
Chromothripsis-like patterns (CTLP)

Input
- 22347 arrays (18458 cases)

Output
- 1566 chromosomes (1028 arrays)
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](image-url) 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).
Progenetix Contributes to ELIXIR hCNV
hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

- Reinforce work on priority areas established in the current hCNV Implementation Study
- Extend collaborations with the Rare Diseases and Galaxy Communities, EJP-RD and GA4GH

**Expected outcomes**

- Shared CNV resources testing advanced versions of the Beacon protocol
- Integration of GA4GH standards such as Phenopackets in such resources
- Tools for data ingestion and export for standard formats (e.g. VCF, Phenopackets) and CNV-specific improvements of such standards
- ELIXIR AAI demo on clinical and research hCNV resources
- Demonstration of Galaxy pipeline adoption for real-world hCNV data analysis projects

- Connecting to international partners, e.g. Cancer Genomics Consortium (U.S.)

- WP1 - hCNV community reference resources
- WP2 - hCNV Resources and Beacon
- WP3 - Galaxy Community Intersection and Data Exchange
- WP4 - Workflows and Tools for hCNV Data Exchange Procedures
- WP5 - Training and dissemination
hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

- WP1 - hCNV community reference resources
- WP2 - hCNV Resources and Beacon
- WP3 - Galaxy Community Intersection and Data Exchange
- WP4 - Workflows and Tools for hCNV Data Exchange Procedures
- WP5 - Training and dissemination

"Galaxify", "Beaconize" & "Phenopack" Progenetix & RD-CNDb prototypes
Beacon API Leads
Jordi Rambla
Anthony Brooks
Juha Törnroos
Discovery WS
Michael Baudis (Beacon)
Marc Fluede (Networks)
ELIXIR
Gary Saunders
David Lloyd
Serena Scollen
Dylan Spalding
Beacon Team CRG
Laureen Fromont
Babita Singh
Sabela de la Torre Pernas
...
Beacon v2 Scouts
Tim Beck
Joaquin Dopazo
Veronique Geoffroy
Jean Muller
David Salgado
Alex Wagner
...
{S}[B] and GA4GH
Melanie Courtot
Helen Parkinson
many more ...

Baudisgroup @ UZH
Michael Baudis
(Paula Carrio Cordo)
(Bo Gao)
Qingyao Huang
Sofia Pfund
Rahel Paloots
Hangjia Zhao
Pierre-Henri Toussaint

Beacon Protocol for Genomic Data Sharing
Beacon provides discovery services for genetic data using the Beacon API developed by GA4GH. The Beacon protocol is a standard for genomics data discovery. It provides a framework for public web services against genomic data collections, for instance from population based or disease specific repositories.

Beacon v2: Towards Flexible Use and Clinical Applications
The original Beacon protocol had been designed to be:
- Weight: focus on robustness and easy implementation
- Federated: maintained by individual organizations and assembled into a collection
- General purpose: used to report on any variant collection
- Aggregatable: provide a breakdown or quantification answer about the data
- Privacy protecting: queries do not return information about single individuals
- Sites offering beacons can scale through aggregation Beacon Networks, which queries joining a potentially large number of international Beacon instances
- Since 2015 the development of the Beacon protocol has been lead by ELIXIR and international participants. Recent versions of the Beacon protocol have extended:
  - Providing a framework for other types of genomic variant data (e.g. variants)
  - Allowing for data delivery using RESTful protocols, e.g. to link with clinical environments and allow for data delivery and visualization services
Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

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