A Reference Resource for Copy Number Variations in Cancer

Implementing GA4GH Standards to Drive an Open Oncogenomics Resource
<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Position</th>
<th>Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>Heidelberg</td>
<td>Student of medicine</td>
<td>Doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter)</td>
</tr>
<tr>
<td>2001</td>
<td>Stanford</td>
<td>Post-doc in hemato-pathology</td>
<td>Molecular mechanisms of leukemogenesis</td>
</tr>
<tr>
<td>2003</td>
<td>Gainesville</td>
<td>Assistant professor in paediatric haematology</td>
<td>Molecular mechanisms of leukemogenesis</td>
</tr>
<tr>
<td>2006</td>
<td>Aachen</td>
<td>Research group leader in genetics</td>
<td>Genomic array analysis for germline alterations</td>
</tr>
<tr>
<td>2007</td>
<td>Zürich</td>
<td>Professor of bioinformatics</td>
<td>Systematic assembly of oncogenomic data</td>
</tr>
</tbody>
</table>
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

---

**Global Alliance for Genomics & Health**
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)
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

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

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))
History & Current State...

Origins & trajectory of the Progenetix Resource
**Comparative Genomic Hybridization**

**Molecular-Cytogenetic Technology for Genomic Imbalance Screening**

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)

- Based on *in situ* suppression hybridization of labeled genomic tumor and reference DNA against a karyotypically normal metaphase chromosomes

- Analysis of relative fluorescence ratio allows semi-quantitative copy number read-out

- Indirect attribution of involved target genes through cytogenetic bands (megabase resolution)

---


---

**Comparative Genomic Hybridization (CGH)**

- Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes in the copy number in the tumor DNA. Michael Baudis, 1998
Comparative Genomic Hybridization
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on in situ suppression hybridization of labeled genomic tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows semi-quantitative copy number read-out
- indirect attribution of involved target genes through cytogenetic bands (megabase resolution)


+6p, -6q
Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets
Let's build a database!

... using archaic tools
Progenetix CGH Database and Website

- originally an internal FileMaker Pro database, to store CGH profiles and annotations for the "Organization of Complex Genomes" group (head: Peter Lichter) 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
- in 2000 online version

- Dec 6, 2000
  - first time online
- Nov 30, 2000
  - addition of graphical representation and gene table
- Nov 17, 2000
  - generation of website layout and database automatisation
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

Clustering in 451 NHL: +12 in CLL and extranodal NHL

Material and Methods

Chromosomal aberration data of more than 5470 cases from 195 publications describing results of Comparative Genomic Hybridization (CGH) experiments were collected. Minor requirements were disposition of a malignancy in human neoplasias, absence of clinical tumor samples and report of the analysis results on a case by case basis, resolved to the level of single chromosomal bands. Data was transformed from the diverse annotation of the original publications into a standardized aberration matrix. For the transformation of the non-linear ISCN data to a two-dimensional matrix of bands, a reverse pattern matching algorithm was developed in Perl. Graphical representations and cluster images are generated for all different subsets (pathologies, ICC-D-2 entities, meta-groups) and presented on the progenetix.net website.

Results

Out of 4606 tumor samples, 3883 (79%) showed chromosomal imbalances for CGH. The average tumor band probability was 0.05 for a base rate of 0.02 and 0.04 for a gain rate of 0.05. Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (19 tumor cases) with the strongest hot spots for gains were small cell lung carcinomas (18% with max. 96.2% at 3p14p36) and metastatic adenocarcinomas (15.9% with max. 95.7% at 11q13). Prominent gain regions were +4q73.2 (4%), +12p21.3 (5.9%), +12q13 (4.7%) with max. 95% at 6q16 to 7q31, T-PLL (4.7% with max. 95% at 9q34 to 10q25) and MZBCL (4.6% with max. 95% at 6q16 to 7q31). By cluster analyses, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsumed in the tumor diagnostic entity, the example of neoplasia is shown.

Conclusion

So far the progenetix.net project was able to:

1. collect a large dataset of genomic aberration data generated through a molecular cytogenetics-screening technique (CGH),
2. develop the software tools to transform these data to a meta format compatible to commonly used genomic interval descriptions,
3. produce graphical and numerical output from those data for hot spot extraction and statistical analysis.

For future approaches, the collection data will be valuable for filtering the data from expression array experiments for relevant genes, and the clustering data can be used for different ways to the oncogenic process of different tumor entities. The transformed raw data of the progenetix.net collection is available for research purposes over the website.

Clustering of the band averages for the different ICC-D-2 online CGH databases in the ‘All lymphoid’ related disease entities hematopoietic and lymphoid tissue neoplasms, melanoma and sarcomas.

Examples of subtypes of genomic imbalances

SCCL, pseudochromosomes, high grade GCCL, T-PLL, differentiated biphenotypic

Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

Michael Baudis, dept. of pathology, stanford university

Michael Baudis | BCATS Biocomputing at Stanford II | Stanford Nov 2002

Michael Baudis | Presentation ASH | Orlando 2001-12-05
Progenetix Database in 2003
Text conversion for CNVs

• based on listed CGH results from publications
  ✷ literature detection using optimized PubMed queries
  ✷ extraction (copy/paste, typing) of revish ISCN karyotypes from articles and supplementary material
  ✷ annotation cleanup using scripting with regular expressions (Perl)
  ✷ custom script to convert cleaned ISCN annotations to cytoband status maps
  ✷ custom graphics libraries to create graphical representations of CNV frequencies
Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies
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)
Array-based Detection of Copy Number Variations

Gain of chromosome arm 13q in colorectal carcinoma

MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

low level/high level copy number alterations (CNAs)

2-event, homozygous deletion in a Glioblastoma
arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer

visualizing cancer genome array data @ arraymap.org

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

The current data reflects:

- 7224 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 Cytoogenetics and Oncogenomics” 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 Baudits regarding a collaborative project.

© 2000 - 2019 Michael Baudits, refreshed 2019-06-12/18:19/00:182 in 6.0.6. on server 100.191.1.120:444, 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

• merging of arrayMap (i.e. probe access enabled) and annotation derived (aCGH, WGS, WES, other arrays) data
• 115’357 cancer CNA profiles
• systematic metadata annotations following GA4GH standards
• unrestricted access w/o registration
• data access API
• online visualization
• CNA statistics
Cancer Type CNA Data

- hierarchical aggregation of cancer samples
- pre-computed CNA frequencies for fast overview
- sample retrieval for custom grouping, visualization

Progenetix

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

- NCIT:C3262: Neoplasm (116232 samples)
- NCIT:C3263: Neoplasm by Site (109317 samples)
- NCIT:C36482: Genitourinary System Neoplasm (16410 samples)
- NCIT:C2910: Breast Neoplasm (15525 samples)
- 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, Perineural, and Schwann Cell Neoplasm (11690 samples)
- NCIT:C4788: Malignant Nervous System Neoplasm (11608 samples)
- NCIT:C38571: Malignant Cranial Nerve Neoplasm (19 samples)
- NCIT:C3716: Primitive Neuroectodermal Tumor (2213 samples)
- NCIT:C4627: Malignant Central Nervous System Neoplasm (9110 samples)
- NCIT:C64827: Malignant Neoplasm of the Meninges (63 samples)
- NCIT:C4717: Anaplastic Ganglioglioma (1 sample)
- NCIT:C4822: Malignant Glioma (5460 samples)
- NCIT:C5114: Malignant Intracranial Neoplasm (3242 samples)
- NCIT:C62332: Central Nervous System Carcinoma (30 samples)
- NCIT:C60967: Click to retrieve samples for NCIT:C7541 (3357 samples)

Retinoblastoma (NCIT:C7541)

Download SVG

Filter subsets ...
Hierarchy Depth: 2 levels

No Selection

- NCIT:C7541: Retinoblastoma (173 samples)
This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. <= ~1Mbp in size). The query is against the Progenetix and arrayMap collections. It can be modified e.g. through changing the position parameters or diagnosis.

<table>
<thead>
<tr>
<th>Reference name</th>
<th>CDKN2A Deletion Example</th>
<th>MYC Duplication</th>
<th>TP53 Del. in Cell Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>G: 9</td>
<td>21500001-21975098</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Cancer Classification(s)**

- NCIT:C3058: Glioblastoma (4358)

**Biosample Type**

**Filters**

**City**

**Filter Logic**

AND

**End (Range or Structural Var.)**

21967753-22500000

Query Beacon
Sample search by CNV
Variants Tab

Assembly: GRCh38  Chro: 9  Start: 21500001-21975098  End: 21967753-22500000  Type: DEL  Filters: NCIT:C3058

<table>
<thead>
<tr>
<th>Int. ID</th>
<th>Digest</th>
<th>Callset</th>
<th>Biosample</th>
<th>Chr.</th>
<th>Ref. Base(s)</th>
<th>Alt. Base(s)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>5bab57b727983b2e00ca09e</td>
<td>9:21548871-21999595:DEL</td>
<td>pgxcs-kftvmlzxy</td>
<td>pgxbs-kftvgk8h</td>
<td>9</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
<tr>
<td>5bab57b727983b2e00cb055</td>
<td>9:21958233-21999595:DEL</td>
<td>pgxcs-kftvmmj5y</td>
<td>pgxbs-kftvgk00</td>
<td>9</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
<tr>
<td>5bab57b727983b2e00cdc18</td>
<td>9:21958233-21999595:DEL</td>
<td>pgxcs-kftvmmj8y</td>
<td>pgxbs-kftvgka5</td>
<td>9</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
<tr>
<td>5bab57b727983b2e00ce2c6</td>
<td>9:21979189-21999595:DEL</td>
<td>pgxcs-kftvmmurl</td>
<td>pgxbs-kftvgkea</td>
<td>9</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
<tr>
<td>5bab57b727983b2e00ce49a</td>
<td>9:21958233-21999595:DEL</td>
<td>pgxcs-kftvmmmb</td>
<td>pgxbs-kftvgkfr</td>
<td>9</td>
<td></td>
<td></td>
<td>DEL</td>
</tr>
</tbody>
</table>
Progenetix Services, Documentation...

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

• API & documentation "progressing" ...

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 lab-separated columnar text file such as produced by array or sequenom CNV software, with an optional metadata header for e.g. patient grouping instructions.

@mbaudis 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 QV).

Such queries will result in a full-whole chromosome CNV events covering the gene of interest, too, should be narrowed by providing e.g. Variant Type and Minimum Size (e.g. 2000000) values.

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

The Progenetix oncogenic resource in 2021

Qingsen Huang, Paula Carrio Cordoba, Bei Ge, Rahul Palotnis, Michael Baudis

[link] doi: https://doi.org/10.1101/2021.01.10.428327

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 1027 individuals from 16 publications. The measured data include copy number variation as well as (in pan) point mutations on relevant genes, e.g. TP53, NF1, ATRX, TERT promoter.

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

arrayMap is Back

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

@mbaudis 2021-02-06: more ...
Bioinformatics & Data Curation - arrayMap data “Pipeline”

- GEO ArrayExpress
  - amDownload
    - Metadata (GEO soft files...)
      - amParseMeta
        - metadata.ods (multi-sample)
          - amUpdateMeta
            - db.metadata
              - db.experiments
- Array Probe Data (log2 tables, .CEL)
  - Metadata Review
    - Selection
      - Entity Mapping
        (ICD-O, Biosample, Individual, clinical...)
  - Array Processing (re-processed)
  - Publications Direct Submissions
    - Manual interpretation of associated information (article, series description...)

Correspondence may also be addressed to Haoyang Cai. Tel: +86 28 85418843; Fax: +86 28 85412571; Email: haoyang.cai@gmail.com

Somatic genomic alterations refer to DNA sequence alterations (CNAs) that alter gene function. They are important subclasses of genomic copy number abnormalities (CNAs) and can be instrumental in improving the assessment of data used for genomic data mining experiments and diagnoses. Data sets included in arrayMap have been upgraded for a more flexible array data visualization, and new developments in arrayMap content and utilities have been incorporated to address technical obstacles.

The SIB Swiss Institute of Bioinformatics' resources: Nucleic Acids Research, 2015, Vol. 43, Database issue 18

© 2015 Springer Nature - All rights reserved

Published online. For recreation, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, please see the Creative Commons Attribution License.

Creative Commons Attribution License

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Progenetix & arrayMap: 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 sets in tutorials

Data sets in the wild
Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```plaintext
'^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 = Immunotype: 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 from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 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 from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 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_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_series_id = GSE7255
```

---

progenet
Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
SAMPLE = GSM174832
SAMPLE_title = 9194
SAMPLE_geo_accession = GSM174832
SAMPLE_status = Public on May 01 2007
SAMPLE_last_update_date = Mar 13 2007
SAMPLE_type = genomic
SAMPLE_source_name_ch1 = Bone marrow with 96% blasts
SAMPLE_taxid_ch1 = 9606
SAMPLE_characteristics_ch1 = Immunotype: 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 from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
SAMPLE_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 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 from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000.
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_country = Netherlands
SAMPLE_series_id = GSE7255
```

```
foreach (grep { !/characteristics_ch\d/ } @in) {
    my ($key, $value) = split('=', $_, 2);
    $key =~ s/[\^]/\W/g;
    if ($key =~ /submission_date/) {
        $sample->{YEAR} = $value;
    } else {
        $sample->{YEAR} =~ s/[\^\d\d\d\d]/\1/g;
    }
}

SAMPLE->{samplekey} = 'AGE';
SAMPLE->{matches} = [ qw( age ) ];
SAMPLE->{reft}, $sample->{retk} } = _grepmeta( $samplekey, $meta );
if ($sample->{ret} =~ /*\(.+)*/i) {
    if ($sample->{ret} =~ /\d+months/ ) {
        $sample->{ret} = 'months';
    }
    $sample->{id} =~ s/[\^\d\d\d\d]/\1/g;
}
SAMPLE->{samplekey} = _normNumber($samplekey->{retv});
if ($sample->{ret} =~ /\d+/ ) { $sample->{samplekey} = 12 };
if ($sample->{samplekey} == 0) { $sample->{samplekey} = 'NA' };
$sample->{samplekey} = sprintf "%2f", $sample->{samplekey} };
```
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_molecule_ch1 = genomic DNA
Sample_extract_protocol_ch1 = Sample DNA Extraction Protocol
Sample_label_protocol_ch1 = NimbleGen Cy5 Sample Labeling Protocol
Sample_label_protocol_ch1 = Other: The DNA was isolated by Qiagen DNeasy Tissue kit according to the manufacturer's recommendations.
Sample_label_ch1 = cy5
Sample_label_protocol_ch1 = Other: Proprietary protocol information available at http://www.nimblegen.com/technology/index.html
Sample_source_name_ch2 = 481
Sample_organism_ch2 = Homo sapiens
Sample_taxid_ch2 = 9606
Sample_characteristics_ch2 = Gender: male
Sample_characteristics_ch2 = Age: 49
Sample_characteristics_ch2 = Tissue: lymph node
Sample_characteristics_ch2 = Disease state: Lymphoma
Sample_characteristics_ch2 = Individual: 481
Sample_characteristics_ch2 = Clinical info: Submitting diagnosis: DLBCL
Sample_characteristics_ch2 = Clinical info: Final microarray diagnosis: ABC DLBCL
Sample_characteristics_ch2 = Clinical info: Follow up years: 10.75
Sample_characteristics_ch2 = Clinical info: Chemotherapy: CHOP-Like Regimen
Sample_characteristics_ch2 = Clinical info: ECOG performance status: 2
Sample_characteristics_ch2 = Clinical info: Stage: 4
Sample_characteristics_ch2 = Clinical info: LDH ratio: 0.82
Sample_characteristics_ch2 = Clinical info: Number of extranodal sites: 1
```

Channel 1 is normal -> Cave value swap!
Gender or "chromosomal sex"?
context indicates years, but if it would be a medulloblastoma...
Unknown way to express "alive"!
Data Curation

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

- correct data is important for any type of scientific analysis
- errors in formats and values can occur during all steps between data acquisition and analysis (numerous "Excelgates"!)
- "meta"-resources and analyses are prone to erroneous data due to varying input formats and lack of source control

➡ always look for batch effects and outliers!
Data Curation - Geolocations

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

The most geo-tagged place on earth is Null Island

A troubleshooting country has been added with an Indeterminate sovereignty class called Null Island (1, 2). It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern Greenwich prime meridian). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)

https://en.wikipedia.org/wiki/Null_Island

Geographic distribution (by corresponding author) of the 118654 genomic area, 36768 chromosomal COI and 42105 whole genome/exome based cancer genome datasets from the 3306 listed publications. Area sizes correspond to the sample numbers reported from a given location.

Michael Szell: The Data Science Process 2

Progenetix publication collection

2020-11-25

2020-11-28
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 ...)
Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3,229 publications reporting 174,530 individual samples in single series from 1 to more than 1,000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

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.
Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.
30,000 patients will have their genome sequenced for rare-disease diagnosis

70,000 genomes (patients + relatives) will be sequenced to help rare disease diagnoses

23,000 cancer patients will have their genome sequenced

50,000 genomes will be sequenced for cancer diagnosis

36,223,000 rare disease patients will have their genome sequenced

83,000,000 genomes will be sequenced for rare disease diagnosis

123,768,000 cancer patients will have their genome sequenced

248,000,000 genomes will be sequenced for cancer diagnosis

* Projected figures, based on current data and known status of genomics initiatives worldwide.

Source: From Op-ed on BioRxiv

Slide provided by Heidi Rehm
The vision: Federation of data
The Global Alliance for Genomics and Health
Making genomic data accessible for research and health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary, Orlando
- May 2018 - Vancouver
- October 2018 - 6th plenary, Basel
- May 2019 - GA4GH Connect, Hinxton
- October 2019 - 7th Plenary, Boston
- October 2020 - Virtual Plenary ...
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 ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.
Global Alliance “Beacon” - Jim Ostell, NCBI, March 7, 2014

Introduction

... I proposed a challenge application for all those wishing to seriously engage in international data sharing for human genomics. ...

1. Provide a public web service
2. Which accepts a query of the form “Do you have any genomes with an “A” at position 100,735 on chromosome 3?”
3. And responds with one of “Yes” or “No” ...

“Beacon” because ... people have been scanning the universe of human research for signs of willing participants in far reaching data sharing, but ... it has remained a dark and quiet place. The hope of this challenge is to 1) trigger the issues blocking groups ... in way that isn’t masked by the ... complexities of the science, fully functional interfaces, and real issues of privacy, and to 2) in short order ... see real beacons of measurable signal ... from at least some sites ... Once your “GABeacon” is shining, you can start to take the next steps to add functionality to it, and finding the other groups ... following their GABeacons.

Utility

Some have argued that this simple example is not “useful” so nobody would build it. Of course it is not the first priority for this application to be scientifically useful. ... intended to provide a low bar for the first step of real ... engagement. ... there is some utility in ... locating a rare allele in your data, ... not zero.

A number of more useful first versions have been suggested.

1. Provide frequencies of all alleles at that point
2. Ask for all alleles seen in a gene region (and more elaborate versions of this)
3. Other more complicated queries

Implementation

1. Specifying the chromosome ... The interface needs to specify the accession.version of a chromosome, or build number...
2. Return values ... right to refuse to answer without it being an error ... DOS attack ... or because ... especially sensitive ...
3. Real time response ... Some sites suggest that it would be necessary to have a “phone home” response ...
A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

YES | NO | \0
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

A Beacon network federates genome variant queries across databases that support the Beacon API.

Here: The variant has been found in few resources, and those are from disease specific collections.
ELIXIR - Making Beacons Biomedical

- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant effects
- Beacons as part of local, secure environments
  - local EGA ...
- Beacon queries as entry for data delivery
  - handover to stream and download using htsget, VCF, EHRs
- Interacting with EHR standards
  - FHIR translations for queries and handover ...
The original GA4GH Beacon implementation (up to v0.3) was conceived as a protocol for sharing the presence/absence of a given, specific, genomic mutation in a set of data (from patients of a given disease or from the population in general). Although with some potential benefit, e.g. in the area of rare disease diagnostics, it was not designed for clinical use but chiefly to foster data sharing by triggering the inquisitiveness of researchers once some data of interest is discovered in another institution. While later extensions of the protocol (v1.0 - v1.n) expanded the query and response options, this did not deviate from the general "existence of variants in resource X" paradigm.

The simplicity and success of the concept has generated the request of making it more powerful, more useful in healthcare environments. The requests include:

- Allowing more informative queries, like filtering by gender or age
- Allowing to trigger the next step in the data access process, e.g. who to contact or which are the data use conditions
- Jumping to another system where the data could be accessed, e.g. if the Beacon is internal to a hospital, to provide the id of the EHR of the patients having the mutation of interest.
- Including annotations about the variants found, among which the expert/clinician conclusion about the pathogenicity of a given mutation in a given individual or its role in producing a given phenotype.

The process

The GA4GH Beacon group started a set of meetings and interviews with GA4GH Driver Projects and with ELIXIR partners in order to determine the scope of the next generation Beacon. The goal was to be useful without breaking the simplicity that made Beacon version 1 successful.

Interviews were conducted with the following GA4GH Driver Projects:

- Autism Speaks
- BRCA Exchange
- CanDIG
- EGA, ENA, EVA
- EuCanCancer
- European Joint Programme - Rare Diseases
- H3Africa
- GEM Japan
- Genomics England
- Matchmaker Exchange
- SVIP/SPHN
- VICC

**Beacon v2 - Clinical Beacon requirements**

Authors: Jordi Rambla, Michael Baudis, Anthony J Brookes, Lauren Fromont, Claudia Vasallo, Aina Jené
ELIXIR Beacon Network

- developed under lead from ELIXIR Finland
- authenticated access w/ ELIXIR AAI
- incremental extension, starting with ELIXIR Beacon resources adhering to the latest specification (contrast to legacy networks)
- service details provided by individual Beacons, using GA4GH service-info
- registration service ➡ integrator throughout ELIXIR Human Data  
  ➡ starting point for "beyond ELIXIR" feature rich federated Beacon services
Beacon Project - Partner Engagement & Next Steps

- Working with **partner communities & projects** on deploying Beacons
  - ELIXIR hCNV Community
  - European Joint Program on Rare Diseases
  - clinical groups & data initiatives (e.g. Andalucia, Cancer Core Europe, SPHN)
  - variant annotation resources, with optional clinical components (e.g. SVIP-O)
- Improving reference implementation and standards / **compliance** testing
- Beacon v2 "fast forward" development
- aligning w/ GA4GH standards, through "request & adopt" => SchemaBlocks \{S\}[B]
- networks **throughout & beyond** ELIXIR
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...)
A "full match" BeaconCnvRequest is a typical scenario for e.g. matching CNVs in which the whole CDR of a gene has been duplicated. Here, both start and end search intervals lie outside of the region of interest. The maximum size of matched CNVs can be limited through the extend of the outer bounds (start[0], end[1]).
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...)
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

Biosamples
- id
- individual_id
- bioterms
- geo_provenance
- ...

Variants
- id
- biosample_id
- reference_name
- start
- end
- variant_type
- alternate_bases
- ...

Beacon Response
- beacon_response
- handover_id

Handover Action
- phenopackets
- VCF
- graphics

Handover
- handover_id
- biosample_ids
- variant_ids
- individual_ids
- ...

Intersect

Authenticate
- elixir

Global Alliance for Genomics & Health
SIB
University of Zurich UZH
elixir
Beacon & GA4GH

- GA4GH has become the major "go to" international organization for the development of data exchange standards and implementation guidelines for genomics & health data

- essential for its success are national partner organizations such as SPHN which have the opportunity to shape the development of GA4GH through contributions, but importantly can benefit through the alignment with international, "cutting edge" standards developments, thereby avoiding duplicate efforts & resource waste

- The early adoption of protocols and standards such as Beacon and Phenopackets drives innovation and efficient data use, both for biomedical researchers and for clinicians

- While the direct benefit of e.g. local Beacon installations may be limited compared to legacy systems, it opens the door for scaled integration with outside systems

- Beacon v2 specifically is being developed with clinical requirements in mind and will cover a broad range of use cases in precision medicine, rare diseases and cancer

- The active participation of SPHN in GA4GH development projects supports a leading position for Swiss biomedical research and personalized health applications
Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon v2 API

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)
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.

"group 3"

"group 4"

WNT
Drivers? Passengers? Markers?
Disentangling CNA Patterns

Thousands of genes involved (passengers)
Descriptive report with arbitrary cutoff

Ductal Breast Carcinoma
Glioblastoma
Intra-Disease CNA patterns are not random

1. Robust

Independent evolvement of CNV landscape

2. Regulated

P53 mutant slightly higher CNV

3. Functionally relevant

Distinctive CNV patterns among IntClust groups

Data source: METABRIC 1992 breast cancer samples, hierarchical clustering, tree not shown
Glioblastoma Multiforme
arrayMap (326)
cBioPortal (547)
TCGA (607)

Colon adenocarcinoma
arrayMap (402)
TCGA (283)
cBioPortal (833)

Clear cell renal cell carcinoma
TCGA (520)
arrayMap (242)

Lung squamous cell carcinoma
arrayMap (358)
TCGA (467)

Bladder urothelial carcinoma
TCGA (343)
cBioPortal (425)

Ovarian serous cystadenocarcinoma
TCGA (567)
arrayMap (324)

Lung adenocarcinoma
arrayMap (607)
TCGA (325)

Ductal breast carcinoma
arrayMap (816)
TCGA (778)

Prostate adenocarcinoma
TCGA (290)
arrayMap (281)
Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles
Unique Patterns of Copy Number Mutations Across Cancer Types

An extensive collection of tumor CNV

- Unique & distinctive patterns
- Patterns of sites and disease
- Identify the origin of a tumor

Examples of unique CNV patterns

**Glioblastoma**

**Medulloblastoma**

**Melanoma**
Progenetix as Example Genomics Resource

Some trajectories ...

- from local database to **online resource**
- from flat database to **hierarchical object storage**
- from dedicated database to mix of **open software tools**
- from static pages to **data driven website**
- from copy, paste, clean to **automated download & process** - still edit & clean
- from registered access to raw data & commercial licensing to **CC BY 4.0** (CC0 for tools)
- from local software development to **open code on Github**
- from standalone resource to federated data, **APIs** and services
Diffuse Intrinsic Pontine Glioma (DIPG) Genomics Repository

The DIPG Genomics Repository is an international collaboration supported by The Cure Starts Now Foundation. It aims to provide a central resource for researchers to investigate genome-wide profiling data from childhood diffuse intrinsic pontine glioma specimens, and additionally for other types of pediatric high grade brain tumors.

This work forms part of a systematic review and meta-analysis of pediatric glioma genomics aimed at collating publicly-available data sets of these diseases in children. In addition, we encourage unpublished or pre-publication data to be submitted to a password-controlled site. We welcome any comments you may have as this resource is being developed.

Michael Baudis, Andre von Büren, Chris Jones
DIPG Genomics Repository Leads

Genomic copy number aberrations in 337 DIPG and pediatric high grade gliomas

2014-04-16: Chris Jones’ group at the ICR involved in recent discovery of ACVR1 as DIPG driver gene
2013-08-19: New "DIPG People" page
More news...
Progenetix now allows generation of topic-specific sites, through use of a "cohort" labeling model.

DIPG/pHGG data is first example for "beyond CNA data", using e.g. EIF4A1 per sample mutation annotations.

serves as model for testing more general expansion of Progenetix as a oncogenomics platform.

currently very, very "beta" ...

Restart DIPG...
Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
**SIB** I 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