Developing Beacons for Data Discovery
Advancing Beacons through *data-driven* implementations

Michael Baudis - #GA4GH2017
Beacon Project
An open web service that tests the willingness of international sites to share genetic data.

> 50 Beacons throughout the world
Developing the GA4GH **Metadata** Schema

- arrayMap for GA4GH
  - metadata schema development through implementation of arrayMap resource data
  - OntologyTerm objects for biodiversity
  - implementation w/ ontology services

Driving **Beacon** Development

- Beacon+
  - CNV/CNA as first type of structural variants
  - disease specific queries
  - quantitative reporting
arrayMap
Resource for copy number variation 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 oncogenic CNV data.

The current data reflects:
- 63,000 genotypic copy number arrays
- 763 experimental series
- 140 array platforms
- 141 ICD-O cancer entities
- 654 publications (PubMed entries)

For the majority of the samples, probe level visualization as well as summarized data representation facilitates 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 Prognosetix project.

arrayMap is developed by the group “Theoretical Cytogenetics and Oncogenetics” at the Institute of Molecular Life Sciences of the University of Zurich.
Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay et al., Cancer Cell 2017, in print)
- Structural variations (DUP, DEL) in addition to SNV
- Diagnosis queries using ontology codes (NCIT, ICD-O)
- Quantitative responses
- Current version uses GA4GH schema compatible database
Beacon Concept

• Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay et al., Cancer Cell 2017, in print)
• Structural variations (DUP, DEL) in addition to SNV
• Diagnosis queries using ontology codes (NCIT, ICD-O)
• Quantitative responses
• Current version uses GA4GH schema compatible database
Beacon range queries allow the definition of a genome region of interest, containing a specified variant or potentially other position related feature.

“fuzzy” matching of region ends essential for inexact features.

current Beacon implementation addresses CNV (<DUP>, <DEL>), as are specified in VCF & GA4GH variant schema.
Beacon+ Concept

- Implementation of cancer beacon prototype, backed by arrayMap and DIPG data set (MacKay et al., Cancer Cell 2017, in print)
- Structural variations (DUP, DEL) in addition to SNV
- Diagnosis queries using ontology codes (NCIT, ICD-O)
- Quantitative responses
- Current version uses GA4GH schema compatible database
GA4GH DWG Metadata Task Team
Wrangling the Schema
biomedical "metadata" in different resources frequently follows incompatible classification systems

medical coding systems are driven by different paradigms compared to biological ontologies (e.g. for cross-species comparisons)

frequently used classifications (ICD, Snomed…) are either not “ontologised” or cannot be referenced in open resources

Federated queries across resources need **curated mappings** of classifications/ontologies
<table>
<thead>
<tr>
<th>ICDM</th>
<th>ICDMORPHOLOGY</th>
<th>NCItcode</th>
<th>NCItlabel</th>
<th>ICDT</th>
<th>ICDTOPOGRAPHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>05210</td>
<td>Carcinoma anaplastic type</td>
<td>C4326</td>
<td>anaplastic oligodendroglioma</td>
<td>C719</td>
<td>brain tumor connective and soft tissue</td>
</tr>
<tr>
<td>90510</td>
<td>Dermatoplastic meningioma</td>
<td>C3647</td>
<td>meningioma</td>
<td>C689</td>
<td>dermatologic and reticuloendothelial systems</td>
</tr>
</tbody>
</table>
| 90700 | Squamous cell carcinoma | C2026 | squamous cell cancer | C140 | skin cancer
| 90800 | Endometrioid adenocarcinoma | C3769 | endometrioid adenocarcinoma | C54 | corpus uteri |
| 90700 | Squamous cell carcinoma | C2026 | squamous cell carcinoma | C44 | skin cancer |
| 90700 | Adenocarcinoma | C2026 | adenocarcinoma | C34 | small intestine excl. rectum and rectosigmoid junction |
| 91700 | Enterothelial type T-cell lymphoma | C3177 | chronic myelogenous leukemia | C42 | hemopoetic and reticuloendothelial systems |
| 98653 | Chronic myeloid leukemia | C3177 | chronic myelogenous leukemia | C42 | hemopoetic and reticuloendothelial systems |
| 98653 | Myeloid leukemia | C3177 | myeloid leukemia | C42 | hemopoetic and reticuloendothelial systems |
| 90803 | Follicular lymphoma grade 3 | C4872 | breast cancer | C50 | breast cancer |
| 90803 | Follicular lymphoma | C4872 | breast cancer | C50 | breast cancer |
| 91403 | Adenocarcinoma | C27745 | lung adenocarcinoma | C809 | unknown |

From 456 pairs of ICD-O terms Morphology and Topography representative of cancer entities in arrayMap:

- Develop Python script to take ICD-O Morphology and Topography labels separately
- QUERY ZOOMA, OxO and OLS to find mapping to NClt

<table>
<thead>
<tr>
<th>NCItcode</th>
<th>NCItlabel</th>
<th>ICDT</th>
<th>ICDTOPOGRAPHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3426</td>
<td>Anaplastic Oligodendroglioma</td>
<td>C719</td>
<td>brain tumor connective and soft tissue</td>
</tr>
<tr>
<td>C3647</td>
<td>Dermatoplastic Meningioma</td>
<td>C689</td>
<td>dermatologic and reticuloendothelial systems</td>
</tr>
<tr>
<td>C2026</td>
<td>Squamous Cell Carcinoma</td>
<td>C140</td>
<td>skin cancer</td>
</tr>
<tr>
<td>C3769</td>
<td>Endometrioid Adenocarcinoma</td>
<td>C54</td>
<td>corpus uteri</td>
</tr>
<tr>
<td>C2026</td>
<td>Squamous Cell Carcinoma</td>
<td>C44</td>
<td>skin cancer</td>
</tr>
<tr>
<td>C2026</td>
<td>Adenocarcinoma</td>
<td>C34</td>
<td>small intestine excl. rectum and rectosigmoid junction</td>
</tr>
<tr>
<td>C3177</td>
<td>Chronic Myelogenous Leukemia</td>
<td>C42</td>
<td>hemopoetic and reticuloendothelial systems</td>
</tr>
<tr>
<td>C4872</td>
<td>Breast Cancer</td>
<td>C50</td>
<td>breast cancer</td>
</tr>
<tr>
<td>C27745</td>
<td>Lung Adenocarcinoma</td>
<td>C809</td>
<td>unknown</td>
</tr>
</tbody>
</table>

Can find mapping to NClt in OxO? Yes
Build OWL file and classify No

- Query Zooma + OLS
- Is mapped to ontology? No
- Is mapped to NClt? Yes
- Query Zooma + OLS
- Is mapped to ontology? No
- Is mapped to NClt? Yes
- MANUAL CURATION of >50%
Beacon+ Concept
Testing Beacons for Data Discovery

- standard Beacon payload (e.g. “exists”)
- testing GA4GH metadata “biocharacteristics” ontology term ids
- multiple datasets can be returned (only one shown here)
- quantitative reporting
- additional information about query & dataset(s)
Study in >1000 rare aggressive childhood brain tumors

157 of those not published previously

Copy number aberration data and selected gene panel represented in DIPG Beacon+

Interface with quantitative returns
Implementing real-world datasets for federated access using GA4GH schema specifications: pHGG

Mackay A, Jones C, Baudis M and many, many others: Integrated molecular meta-analysis of 1,000 pediatric high grade and diffuse intrinsic pontine glioma (2017, Cancer Cell, in press)
Check it Out!

- managed, participation driven projects living on Github: **ga4gh**
- **beacon.arraymap.org**
- **dipg.progenetix.org**
- test datasets & code available through our **progenetix** repositories
  - test
  - comment
  - suggest
  - propose
  - complain …