

Qingyao Huang, Paula Carrio-Cordo, Rahel Paloots, Bo Gao and Michael Baudis

Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland



## The Progenetix Oncogenomics Resource

The Progenetix oncogenomics resource provides sample-specific cancer genome profiling data and biomedical annotations as well as provenance data from thousands of individual cancer studies.

With currently 111'840 sample specific curated genomic copy number number (CNV) profiles from 1600 studies, representing over 780 cancer types (according to NCIT neoplasm core), Progenetix empowers aggregate and comparative analyses which vastly exceed individual studies or single diagnostic concepts.

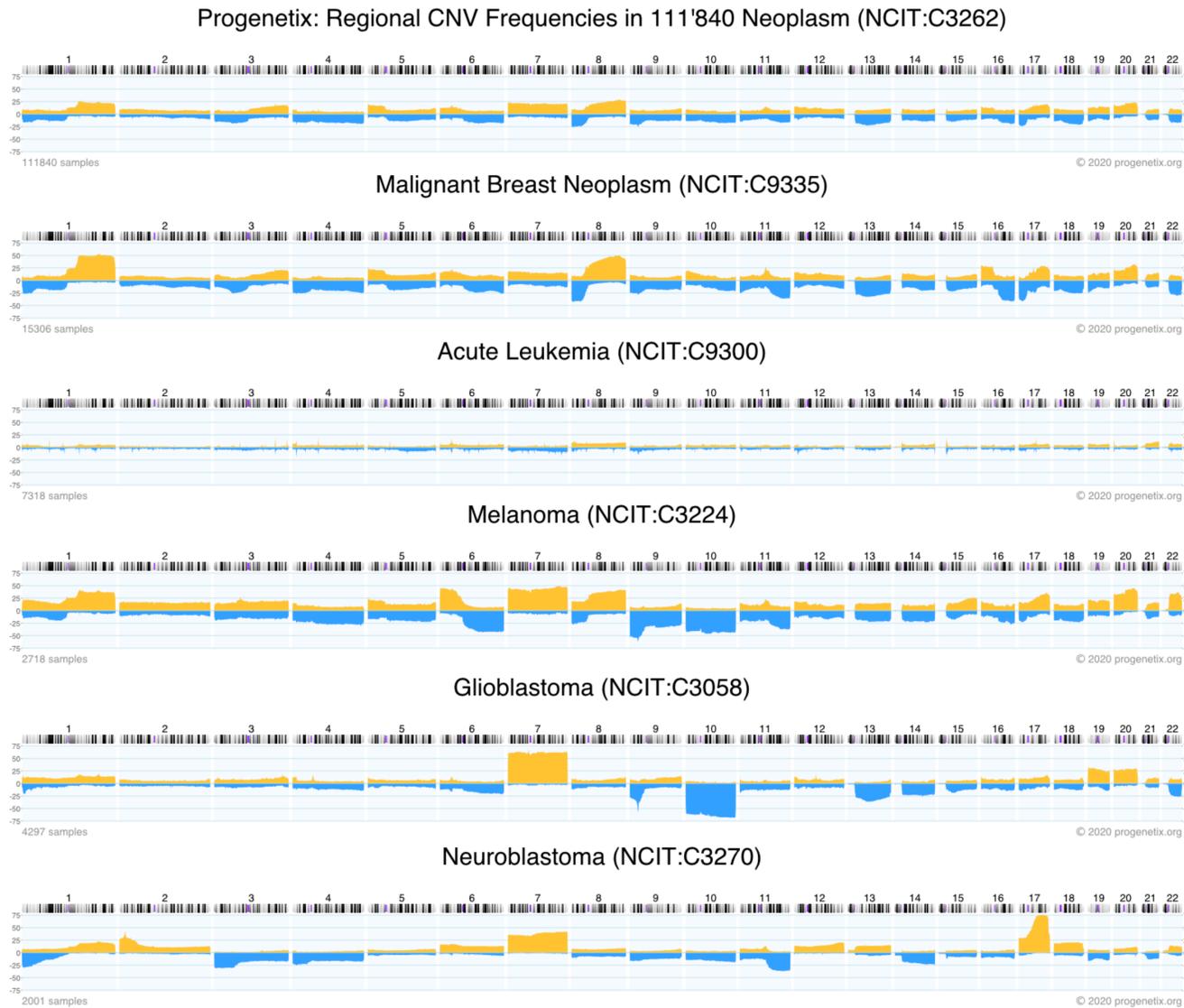
Progenetix provides a demonstration of how an open genomic reference resource can be built around GA4GH standards and how it can be used to support ongoing and future initiatives in GA4GH standard development as well as in driving ELIXIR implementation studies.

## Beacon+ tests Beacon API extensions

The screenshot shows the Beacon+ search interface with various filters. A red box highlights the 'CURIes for Beacon v2 filters' section, which includes 'Gene Spans', 'Cytoband(s)', 'Dataset', 'Genome Assembly', 'Reference name', 'Start or Position', 'End (Range or Structural Var.)', and 'Cancer Classification(s)'. The 'Cancer Classification(s)' dropdown is set to 'NCIT:C3058: Glioblastoma (219) X'.

Beacon+ - built on top of the Progenetix infrastructure - has been instrumental in developing and testing Beacon extensions such as structural variant queries and handover data delivery (v1.n) or filters for querying biological and technical annotations (v2.n).

## Progenetix provides regional CNV frequency profiled for most cancer types



Genomic copy number frequency profiles in some tumor types, plotted from the Progenetix database API. The histograms detail the frequency of genomic duplications/amplifications (yellow; up) and deletions (blue, down) for the corresponding region in a given tumor type or all samples (top).

## Modern Hierarchical Ontologies for Flexible Data Use

The use of hierarchical ontologies for biological classifications and parameters as well as for identifiers and technical metadata is imperative to make data accessible, reusable and amenable to computational mining and analysis methods.

In Progenetix the systematic integration of "classical" property codes (e.g. International Classification of Diseases in Oncology; ICD-O 3) and their translation into hierarchical ontologies with registered identifiers (e.g. NCIT Neoplasm Core, MONDO, EFO...) empowers internal data structures as well as federated query implementations such as through Beacon v2 "filters".

The screenshot shows the 'Cancer Types' filter interface. It features a 'Cancer Classification' dropdown menu with options like 'NCIT', 'ICD-O Histo', and 'ICD-O Topo'. Below the dropdown is a search bar and buttons for 'Filter cancer...', 'Expand', and '3 levels'. A tree view shows the following hierarchy:
 

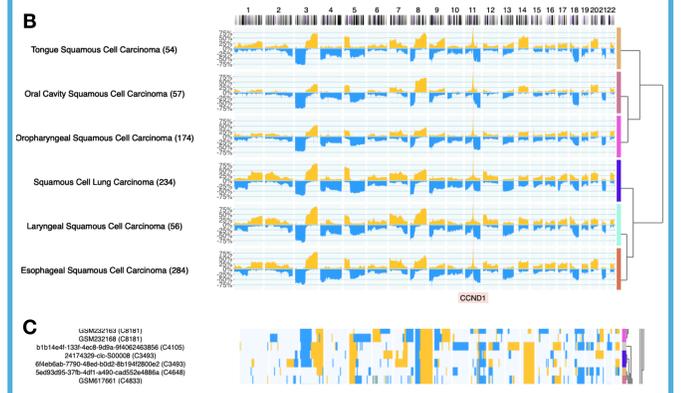
- NCIT:C3262: Neoplasm (111840 samples)
  - NCIT:C3263: Neoplasm by Site (106563 samples)
    - NCIT:C156482: Genitourinary System Neoplasm (16309 samples)
    - NCIT:C2910: Breast Neoplasm (15334 samples)
      - NCIT:C27939: Lobular Neoplasia (92 samples)
      - NCIT:C36083: Intraductal Breast Neoplasm (275 samples)
        - NCIT:C27942: Intraductal Proliferative Lesion of the Breast (270 samples)
        - NCIT:C36090: Intraductal Papillary Breast Neoplasm (5 samples)
        - NCIT:C40405: Breast Fibroepithelial Neoplasm (41 samples)

## Beacon Handover for Data Analytics

The Beacon protocol omits direct data delivery to avoid potential exposure of patient data. Within this paradigm, the Beacon v1.n handover extension adds options for retrieval or analysis of the matched data, outside the Beacon response itself. Beacon v2 will add other delivery methods to be used in conjunction with authentication and authorization technologies.

The screenshot shows a Beacon query interface. It displays a query for 'Assembly: GRCh38 Chro: 11 Start: 65000000-69641313 End: 69651281-74000000 Type: DUP Filters: NCIT:C2929'. Below the query, there are buttons for 'Calls Variants', 'UCSC region', 'Legacy Interface', and 'JSON Response'. A 'Beacon query including v2 filters' box is highlighted. Below this, there are 'Results' tabs for 'Biosamples', 'Biosamples Map', and 'Variants'. A 'Visualization and download from handover objects' section is also visible.

Standard Beacon queries together with dedicated handover objects in Progenetix enable a variety of services: variant display in the UCSC browser, sample data download (above, A); retrieval sample of specific genome profiles for subgroup visualization (below, B) or single sample similarity comparisons (C; all plots showing CNV data).



## Conclusion

Development of protocols, schemas and standards in biomedical research and genomics applications happens in the continuous interplay between technological advances and epistemic advances.

The agile development principles and extensive data content of the Progenetix resource utilizes and promotes standards of the GA4GH ecosystem.