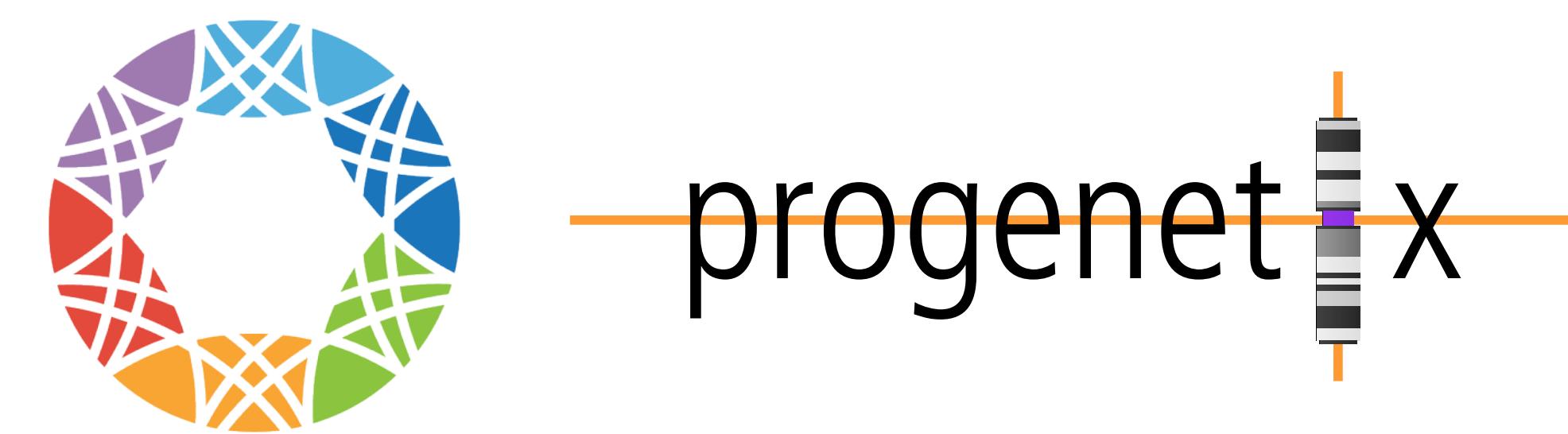


A cancer genomics reference resource and implementation toolkit around GA4GH standards

Qingyao Huang, Rahel Paloots, Hangjia Zhao, Ziying Yang, Paula Carrio-Cordo, 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 for cancer studies. Especially through more than 100k genomic copy number (CNV) profiles from over 500 cancer types, Progenetix empowers comparative analyses vastly exceeding individual studies and diagnostic concepts.

Progenetix has been used in research studies, clinical diagnostics and in the development of data standards for the Global Alliance for Genomics and Health (GA4GH) and the European bioinformatics initiative ELIXIR. The resource's focus on structural genome variants has been instrumental in addressing their specific requirements in GA4GH schema development and the Beacon protocol.

Database URL progenetix.org

License CC-BY 4.0 (CC0 for code)

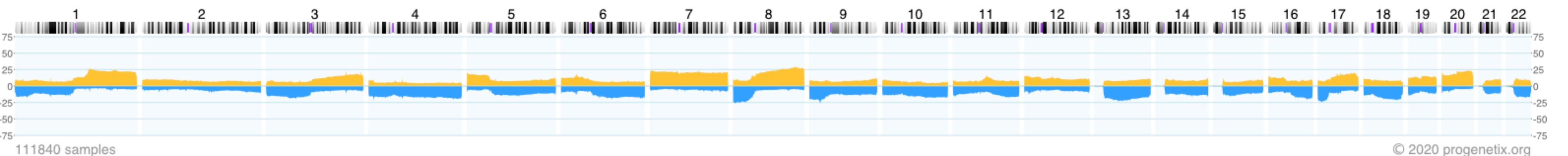
Beacon+ drives Beacon v2 API development

This section shows the Beacon+ interface. It includes a search bar for samples, dropdown menus for CNV Request, Allele Request, Range Query, and All Fields. A 'CNV Example' box contains a query for CNV deletion variants overlapping the CDKN2A gene's coding region. Below it, another box shows a query for copy number queries ("variantCNVrequest"). A 'CURIEs for Beacon v2 "filters"' box highlights the 'Cancer Classification(s)' field, which is set to 'NCIT:C3058: Glioblastoma (2119)'. Other filter options like 'Gene Spans' and 'Cytoband(s)' are also visible.

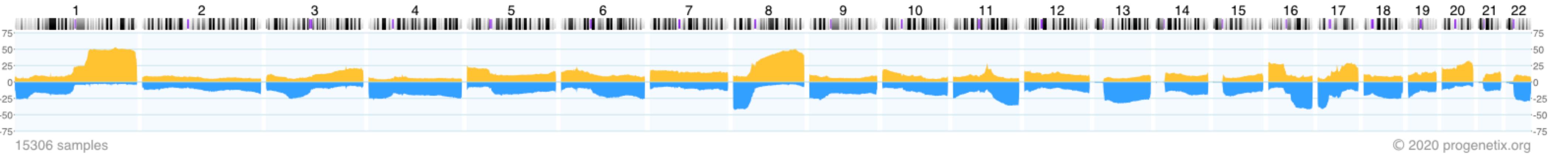
Beacon+ - built on top of the Progenetix infrastructure - has been instrumental in developing and testing Beacon extensions such as **structural variant** queries and **handover** data delivery (v1.n) or **filters** for querying biological and technical annotations (v2.n).

Progenetix provides regional CNV frequency profiled for most cancer types

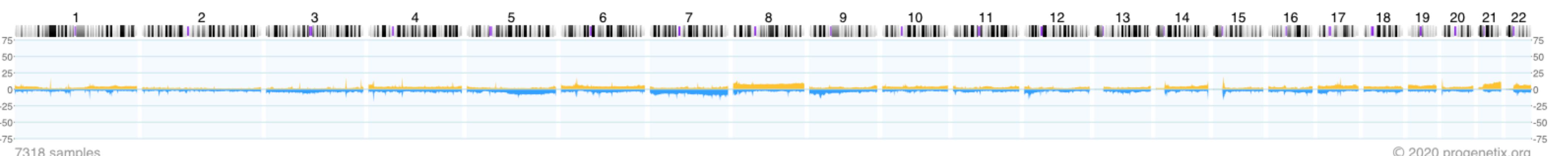
Progenetix: Regional CNV Frequencies in 111'840 Neoplasm (NCIT:C3262)



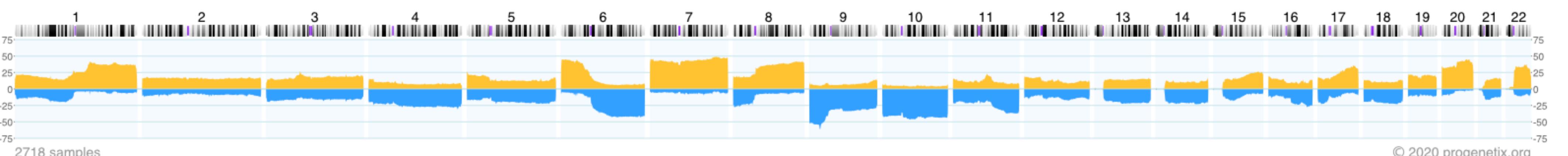
Malignant Breast Neoplasm (NCIT:C9335)



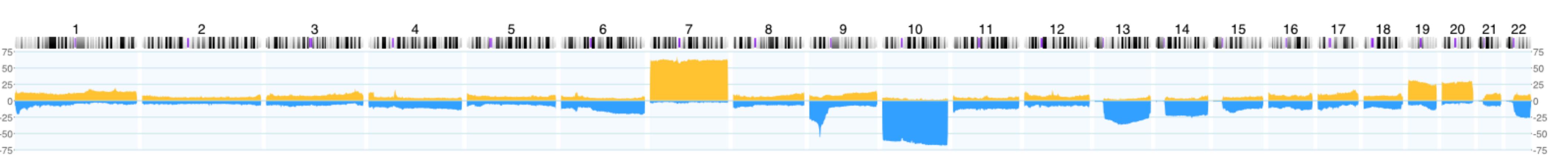
Acute Leukemia (NCIT:C9300)



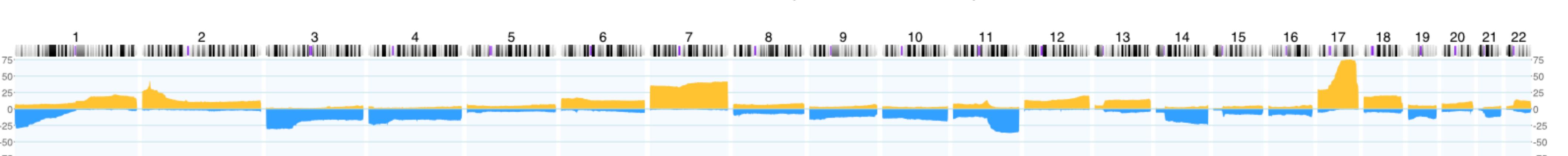
Melanoma (NCIT:C3224)



Glioblastoma (NCIT:C3058)



Neuroblastoma (NCIT:C3270)



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

During development of GA4GH metadata concepts and schemas - which influenced standards such as the Phenopackets format - cancer specific annotations from Progenetix have informed conceptual requirements and domain-specific mappings.

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".

This section shows the 'Cancer Types' interface. It features a hierarchical tree view under 'Cancer Classification'. Nodes include 'NCIT' (selected), 'ICD-O Histo', 'ICD-O Topo', and a 'Dataset' dropdown set to 'progenetix'. A 'Filter cancer...' button and a 'No Selection' button are also present. The tree structure shows 'NCIT:C3262: Neoplasm (111840 samples)', 'NCIT:C3263: Neoplasm by Site (106563 samples)', and several sub-nodes for different cancer types like 'Genitourinary System Neoplasm' (16309 samples) and 'Breast Neoplasm' (15334 samples).

This section shows a detailed view of cancer types. It lists 'NCIT:C3262: Neoplasm (111840 samples)', 'NCIT:C3263: Neoplasm by Site (106563 samples)', and other nodes like '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)', and 'NCIT:C40405: Breast Fibroepithelial Neoplasm (41 samples)'. A 'Dataset' dropdown is set to 'progenetix'.

Progenetix Data API

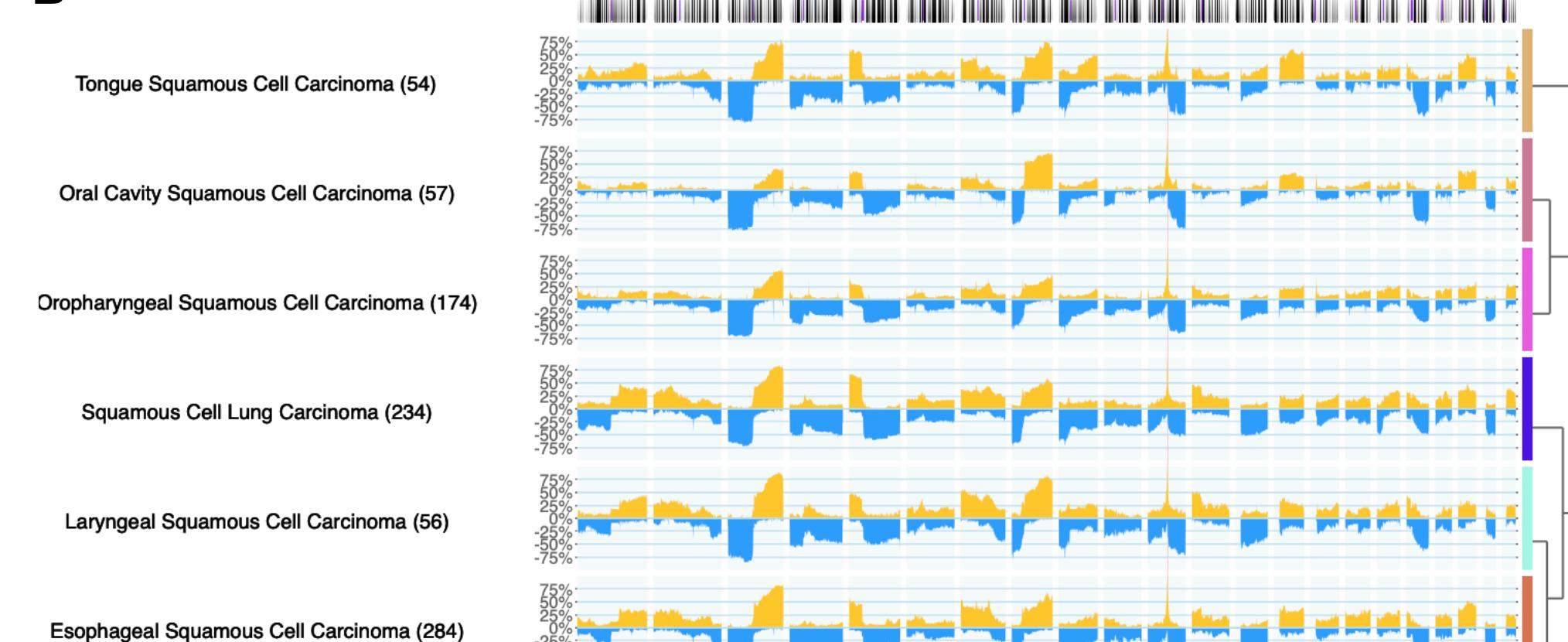
The Progenetix API provides download and data visualization options, with standardized JSON responses and data handling facilitated through extensions like the **pgxRpi** library.

A

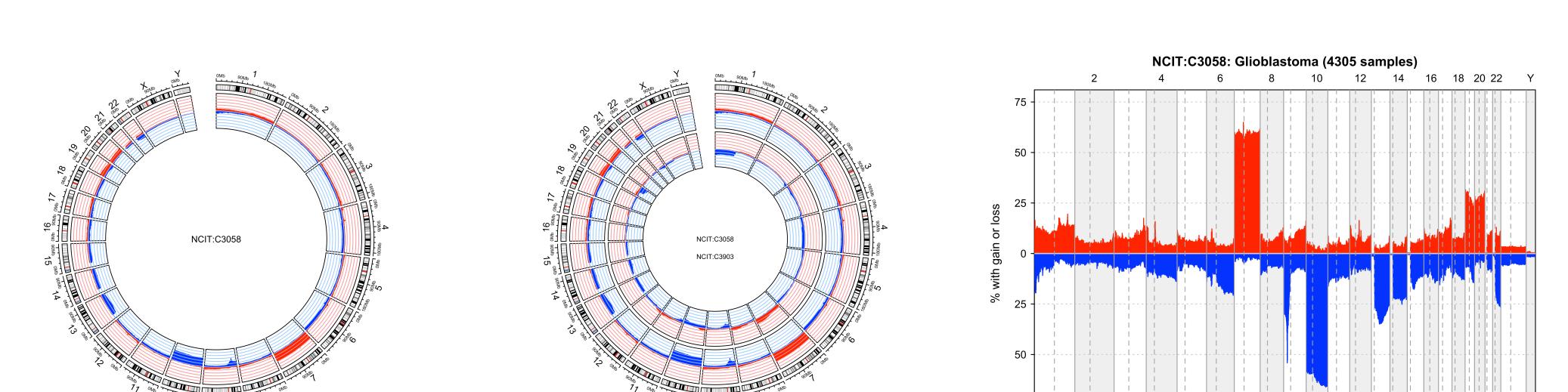
Assembly: GRCh38 Chro: 11 Start: 65000000-69641313 End: 69651281-74000000 Type: DUP Filters: NCIT:C2929

This section shows the Progenetix Data API interface. It displays a histogram of genomic duplication (yellow) and deletion (blue) frequencies for GRCh38 chromosome 11. The y-axis ranges from -75 to 75. The plot is titled '© 2020 progenetix.org'. Below the plot, there are sections for 'progenetix' (Samples: 991, Variants: 894, Calls: 995, Falleles: 0.000123), 'Calsets Variants' (link), 'Legacy Interface' (link), 'Variants in UCSC' (link), 'UCSC region' (link), 'JSON Response' (link), and 'Visualization and download from handover objects' (link). Buttons for 'Results', 'Biosamples', 'Biosamples Map', and 'Variants' are shown. A note says 'See more visualization options.'

B



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); retrieval sample of specific genome profiles for subgroup visualization (B).



Conclusion

We demonstrate how an open genomic reference resource has been built around emerging GA4GH standards and how it is being used to support ongoing and future developments in GA4GH and ELIXIR implementation studies, including an introduction about utilizing the Progenetix code repositories for genomics resource development.

