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

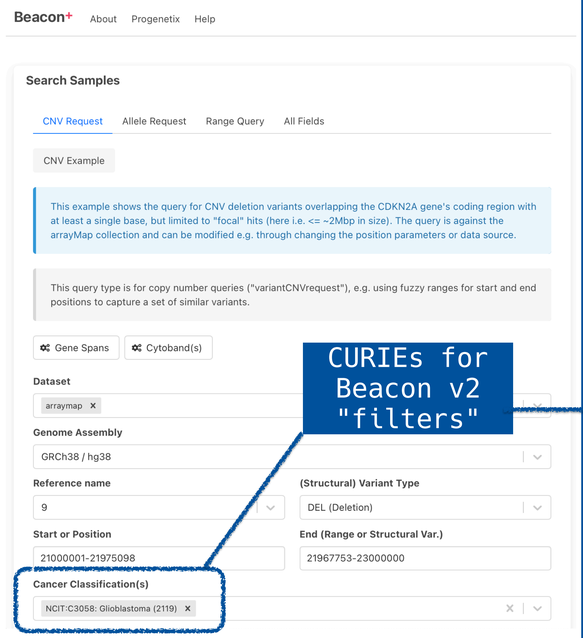
The Progenetix oncogenomics resource provides sample-specific cancer genome profiling data and biomedical annotations as well as provenance data for cancer studies. Especially through more than 100k genomic copy number number (CNV) profiles from over 500 cancer types, Progenetix empowers comparative 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

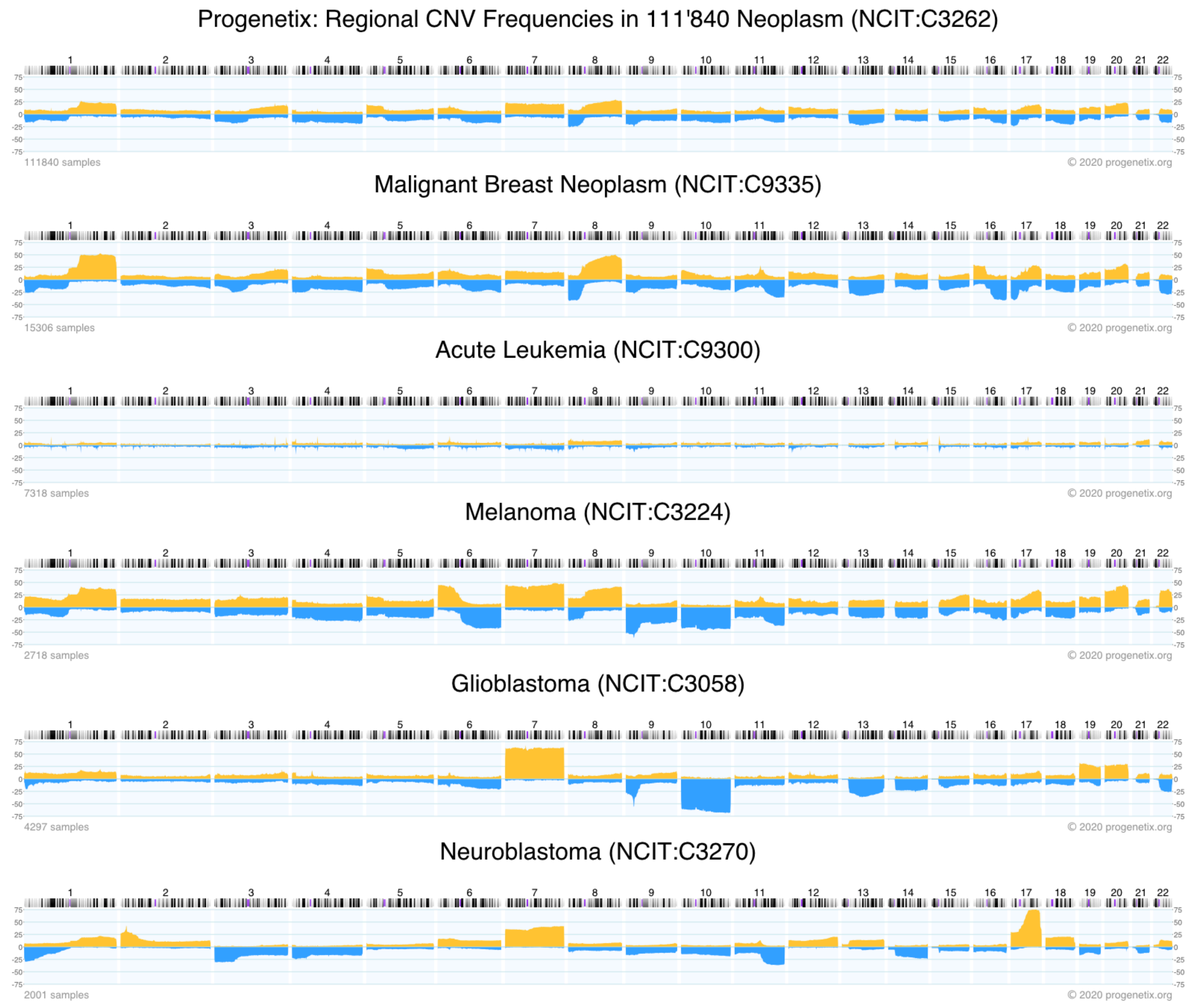
License CC-BY 4.0 (CC0 for code)

Beacon+ drives Beacon v2 API development



*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

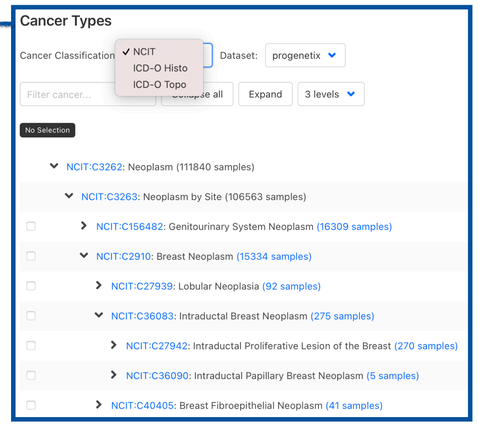


Genomic copy number frequency profiles in some tumor types, plotted from the Progenetix database API. The histograms detail the frequency of genomic duplications/ amplifications (yellow; up) and deletions (blue, down) for the corresponding region in a given tumor type or all samples (top).

Modern Hierarchical Ontologies for Flexible Data Use

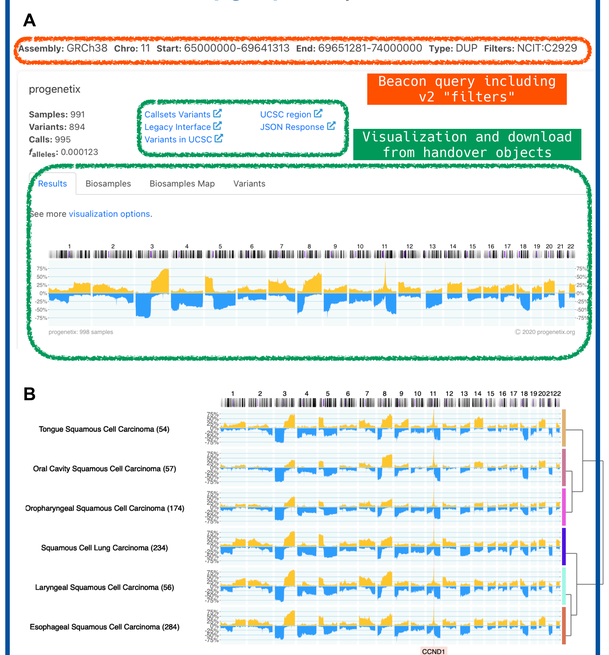
During development of GA4GH metadata concepts and schemas - which influenced standards such as the Phenopackets format - cancer specific annotations from Progenetix have informed conceptual requirements and domain-specific mappings.

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".

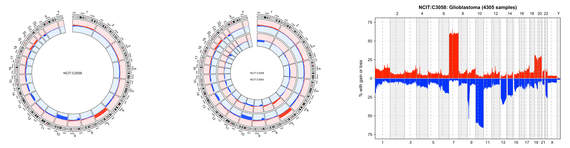


Progenetix Data API

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



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).



Visualization of CNV features using the recently introduced pgxRpi R package. Here, aggregated CNV data for single or two (center) cancer types using Circos or frequency plots in a local R environment.

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.

For more information please visit progenetix.org and the code repositories at github.com/progenetix

Michael Baudis 2021-07-21