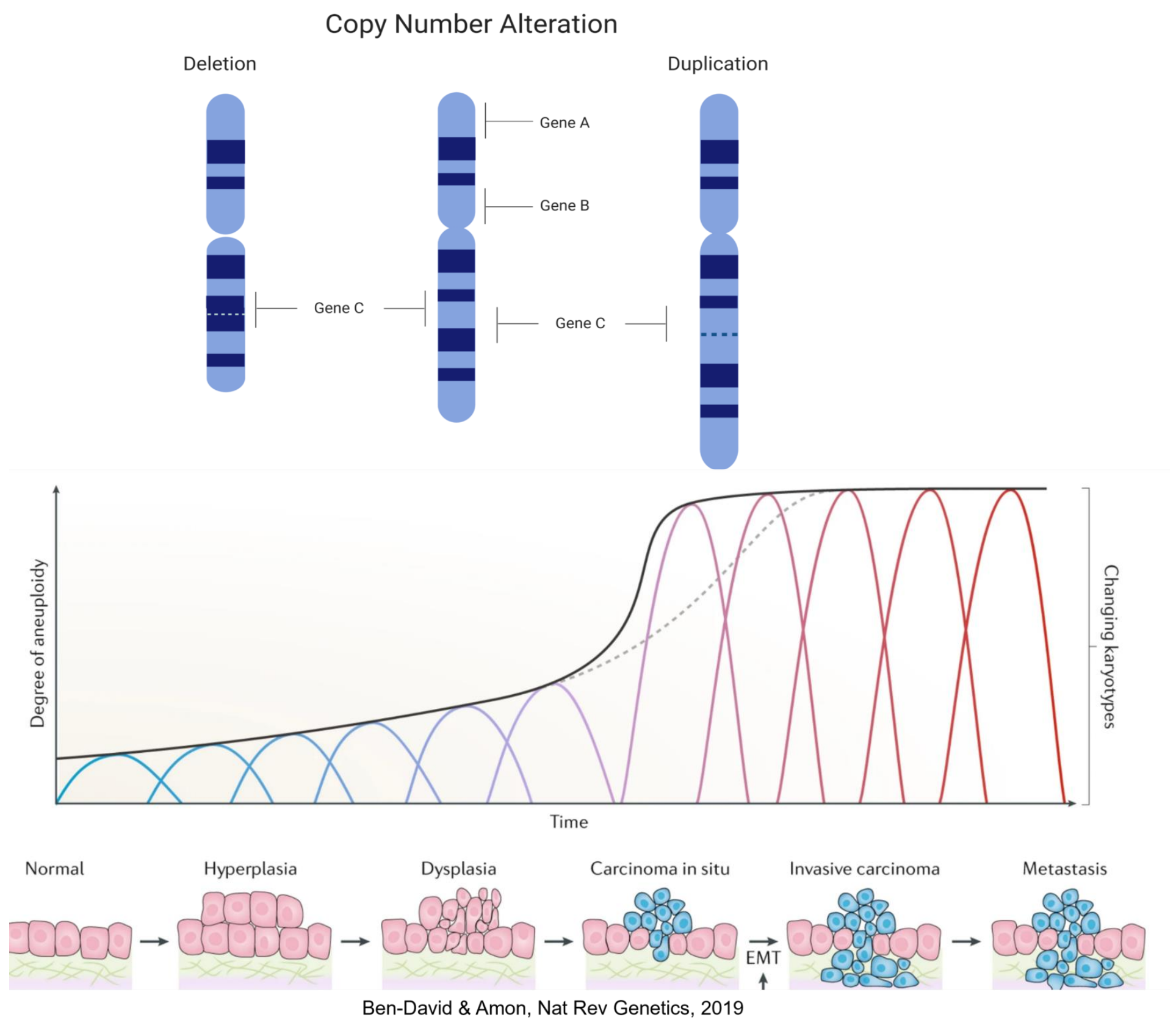
An Open Pan-Cancer Copy-Number-Profile Database From Published Array-Based Studies

Antonia Vlaicu [1], Maxime Tarabichi [1]
1- IRIBHM - Université Libre de Bruxelles (ULB)

Introduction

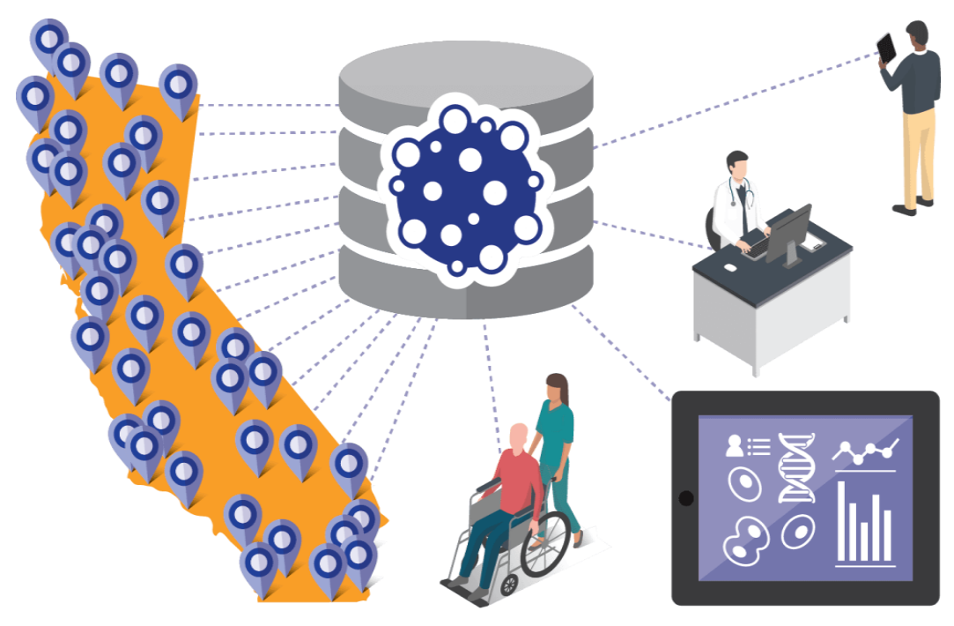
Copy Number Aberrations (CNAs), such as **gains/amplifications**, **losses/deep deletions**, and **whole genome doubling**, are one of the most important type of oncogenic variations. CNAs can activate oncogenes by amplifying them and deactivate tumor suppressor genes through homozygous deletions. [1] They are **crucial in cancer progression,** and chromosomal instability **(CIN)** leading to CNAs provides an important substrate for transcriptomic plasticity through gene dosage effects.

Hundreds of studies on cancer patients have already been conducted using **high-resolution microarrays**, and the genomic data resulting from these studies is publicly available on databases such as Gene Expression Omnibus. These valuable resources, along with **copy-number variant calling bioinformatic tools**, can thus be used to infer the copy number profiles (CNPs) of thousands of cancer patients, analyze them and provide an easy way to access them in the form of an **open database** for the study of CIN.



Ben-David & Amon, Nat Rev Genetics, 2019

Build and publish a database of curated high-quality copy-number profiles



Adding to already-existing **CNP databases,** our database could help the scientific community to :

- Analyze key copy-number variations
- Identify signatures of CNAs across cancer types
- Determine driver genomic regions
- Improve clinical diagnosis

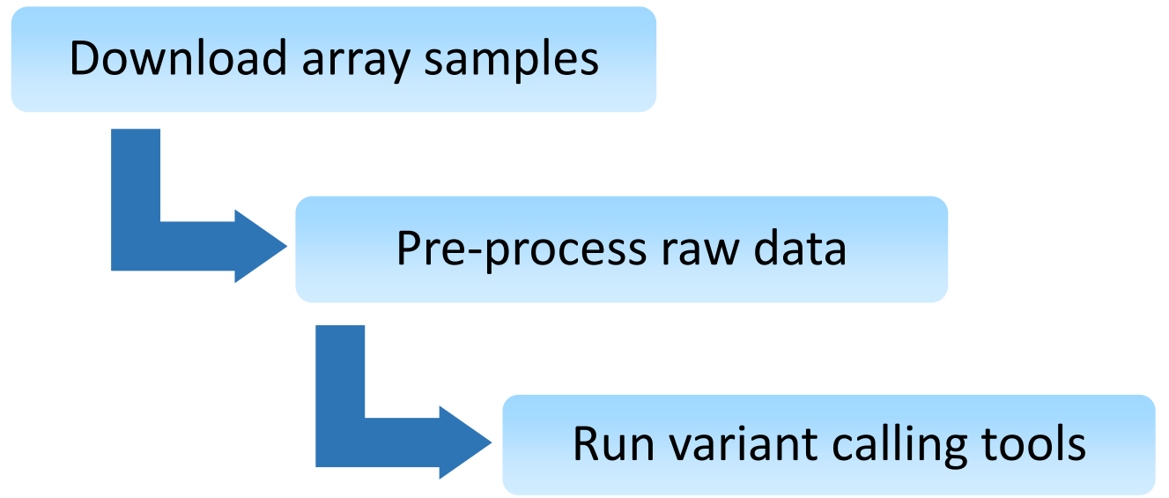
Materials and Methods

1. Sample Retrieval

Gene Expression Omnibus

The first step of the project consists in retrieving the raw data from which the copy number profiles are inferred. We use **high-density SNP and methylation human microarray samples** from cancer related studies. The chosen array platforms are **Affymetrix SNP6, Illumina 450K BeadChip and Illumina Infinium EPICv1**, which cover between **450 000** and **1 million probes**. **Gene Expression Omnibus** provides a total of **18 000 samples** that match our criteria.

2. Copy Number Calling



Download array samples

Pre-process raw data

Run variant calling tools

The second step is the variant calling., for which we used two novel bioinformatic tools: **ASCAT**[2] and **ASCAT.sc**[3]. **ASCAT** infers **allele-specific copy number profiles** from WGS/WES and **SNP array** data. **ASCAT.sc** infers the **total copy number**, and it was designed to work on platforms not covered by ASCAT, including single-cell (sc), shallow-coverage (sc), and targeted sequencing, as well as **methylation arrays**.

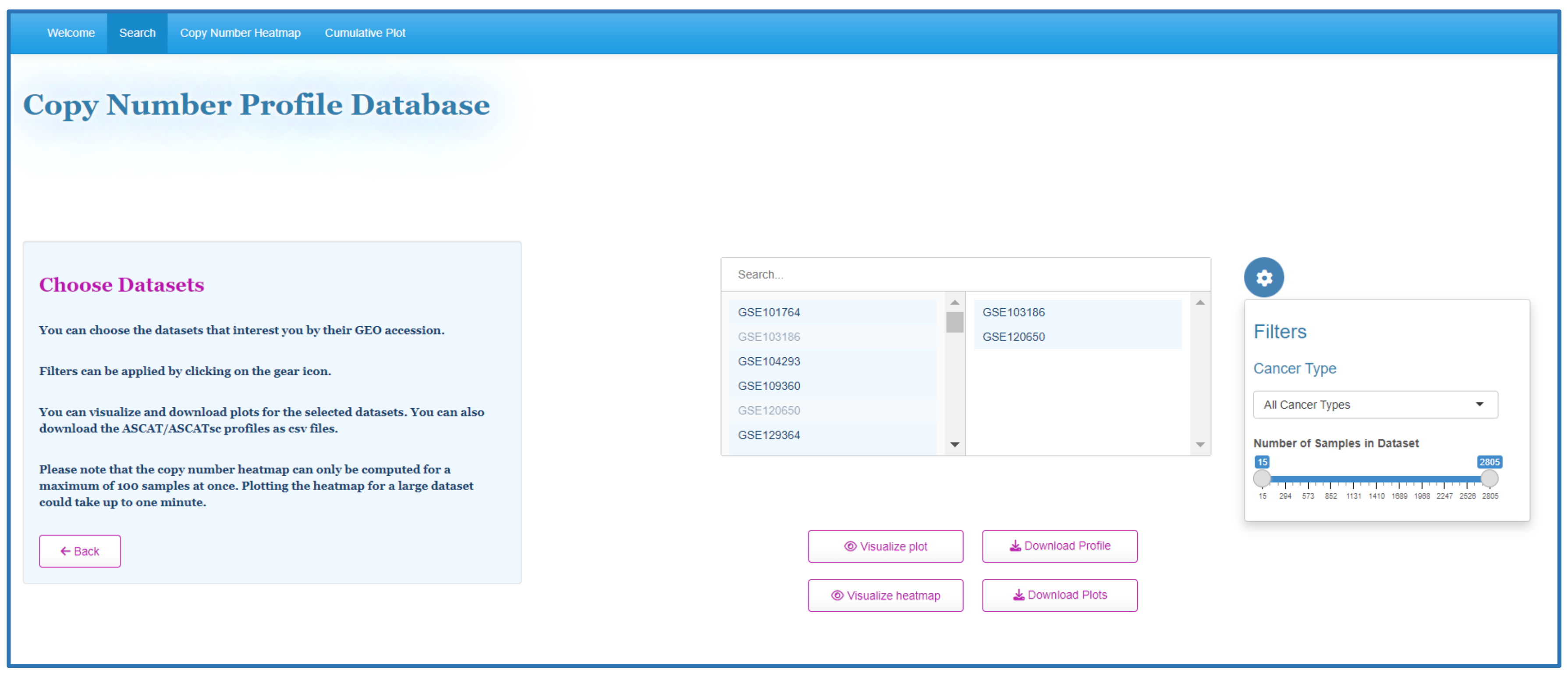
In order to allow for rapid processing of thousands of samples, we implemented **two automated bioinformatic pipelines**, for ASCAT and ASCAT.sc, respectively. The pipelines only require the names and platforms of the desired GEO datasets as input, which are used to download the samples for each dataset, process them and run ASCAT/ASCAT.sc in order to obtain the copy number profiles.

3. Database Construction

Finally, we constructed a database in the form of a **Shiny web application**. Shiny[4] is an R package for the design of biologically oriented, **easy to access applications**. The database contains the copy number profiles, which are compiled in text files. It also allows for quick visualization of the CNPs through the use of **interactive plots**.

Results

The database provides the copy number profiles of 18 000 samples. The CNPs can be browsed per dataset, using the **GEO dataset accession number**. The datasets can be **filtered** according to **the cancer type** or **the number of samples** they contain. Additional **interactive plots** help the user visualize the CNPs of the selected datasets.



Copy Number Profile Database

Choose Datasets

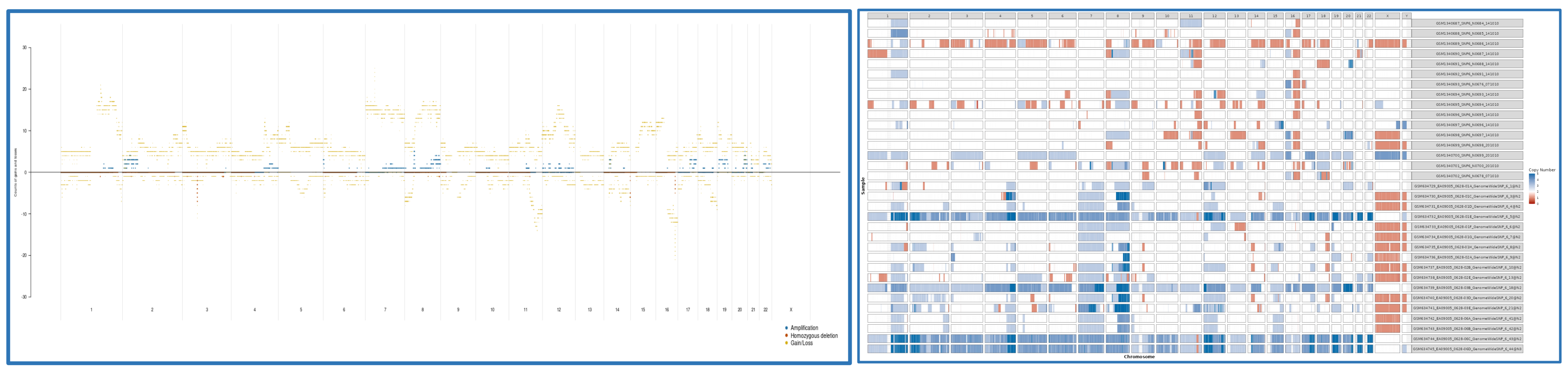
You can choose the datasets that interest you by their GEO accession.

Filters can be applied by clicking on the gear icon.

You can visualize and download plots for the selected datasets. You can also download the ASCAT/ASCATsc profiles as csv files.

Please note that the copy number heatmap can only be computed for a maximum of 100 samples at once. Plotting the heatmap for a large dataset could take up to one minute.

The web application menu allows the user to browse and choose the desired datasets



The user can visualize two types of information: a cross-sample cumulative copy number plot (left) and a per sample copy number heatmap (right). The heatmap shows the profile of each sample of the selected datasets. One sample is represented per line, with the genomic position on the x axis. The color code shows the amplified regions in blue, and the losses or deletions in shades of red. The cumulative plot shows the number of samples out of the total samples that present a gain or a loss for each position in the genome. The amplifications, gains, losses and deletions are color coded (see legend for color key).

Future work

- ➢ Future improvements to the variant calling tools ASCAT and ASCAT.sc to incorporate more platforms would increase the number of samples that they can be run on, and consequently increase the size of the database by thousands of CNPs
- ➢ Additional microarray raw data resources available on other databases, such as ArrayExpress, could also be used to infer CNPs from hundreds of new databases
- ➢ For the database to be fully functional online, it requires a powerful server. In the future, we will make all the functionalities of the database available, as well as use it to help identify signatures across cancer types [5 6]

References

1 - Bo Gao and Michael Baudis. “Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes”. In: Frontiers in Genetics (2021)
2 - Peter Van Loo et al. “Allele-specific copy number analysis of tumors”. In: PNAS (2010)
3 - ASCAT.sc - Copy number from DNA profiling techniques, including single-cell (sc), shallow-coverage (sc), and targeted sequencing, as well as methylation arrays. url: https:/ github .com/VanLoo-lab/ASCAT.Sc
4 - Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023). *shiny: Web Application Framework for R.*, https://github.com/rstudio/shiny, https://shiny.posit.co/.
5 - Steele, C.D., Abbasi, A., Islam, S.M.A. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022)
6 - Drews, R.M., Hernando, B., Tarabichi, M. *et al.* A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022)

FREEDOM TO RESEARCH