Screening for somatic mutations in cancer has become integral to diagnostic as well as target identification procedures. arrayMap is a curated oncomomic resource, focusing on genomic arrays and copy number aberration (CNA) profiles. The underlying data has been extracted from NCBI’s Gene Expression Omnibus (GEO), EB’s ArrayExpress, and, importantly, through targeted mining of publication data. Since its first release in 2012, arrayMap underwent improvements to facilitate meta-analysis of cancer related genome data and clinical use, such as the diagnostic validations as well as target evaluation for personalized therapeutic approaches.

For the 2017 arrayMap update, we expanded both the scope and depth of the database, as well as improved the metadata structure. In a systematic mining of genomic array data from GEO, we evaluated more than 120,000 annotation files. It yielded around 22,000 additional data sets related to somatic mutations in cancer associated samples (cancer specimen or associated reference profiles). Additionally, we were able to increase the database of publications assessed for describing original cancer genome profiling experiments to now more than 3,000 individual articles. The result is a comprehensive and useful database containing information for approximately 40 cancer types and 65,000 genomic array profiles.

Automated patient data analysis

The following data have been retrieved with automatic processes of text analysis:
- Ethnicity (0.5% of the samples have this information),
- Follow-up (0.5% of the samples have this information),
- TNM stage (1% of the samples have this information),
- Death Status (2% of the samples have this information),
- Tumor Grade (2.2% of the samples have this information) and
- Tumor Stage (15% of the samples have this information).

Sex (32% have data)

Based on parsing of the metadata files, male and female sample-based experiments are represented in equal proportions. However, this does not account for samples with sample-specific sex and no annotation in e.g., all prostate samples are male, all ovary samples and most breast samples female.

Age (45% have data)

For every sample, we have parsed manually the anatomical site of origin of the neoplasm (ICD TOPOGRAPHY), and the characteristic of the tumor itself, including its cell type and biological activity (ICD MORPHOLOGY). Here we represent the first 70 entries for ICD TOPOGRAPHY and ICD MORPHOLOGY.

Furthermore, the hierarchical representation of the data in individuals, biosamples and experiments and its association to metadata, such as geographic location of the study, allows for further meta-analysis. In ongoing studies, we will use our data collections to investigate temporal and geographic trends and biases in cancer genome research, to provide a clear view of the current oncomomic research landscape and to identify knowledge gaps, to guide the direction of future studies.

We manually annotated 55,573 individuals and 68,062 experiments. The assignment of the biosample is still in progress.

Manually curated analysis of the study

We have manually curated data for all the samples:

When the study was performed

Source of the sample

Number of samples by technique and country

Channel

ArrayMap & Progenetix 2017: Oncogenic Databases towards Personalized Medicine
Paula Carrio Cordó, Alessio Milanese and Michael Baudis
Institute of Molecular Life Sciences, University of Zurich, Switzerland
Swiss Institute of Bioinformatics, Switzerland