RELEASE NOTES

Message from the project coordinator, Arkaitz Carracedo:

CANCERTOOL was developed to help with the visualisation and analysis of gene expression cancer datasets from international research groups and consortia. From the original release of our tool, modifications and updates will be required, as illustrated by the following examples:

- *The phenotypic (clinical and pathological) and molecular information of the datasets included in CANCERTOOL might be updated in the source repositories.*
- *The guidelines for gene nomenclature and data normalization will be refined by the field.*

We perform a series of test to corroborate that the modifications introduced in new releases do not affect the overall performance of the tool. We have validated that the modifications introduced in latest release have a minor effect on the overall transcriptional landscape of the Datasets. However, the impact on specific genes could be significant. User experience might reveal bugs and errors on the website that need amendment. In this sense, we appreciate all feedback from CANCERTOOL users.

Our aim is to continue to refine and update CANCERTOOL, so that it helps cancer researchers in the discovery and validation of gene expression alterations in human cancer. Multiple additional resources are continuously developed for similar analyses, and we encourage users to exploit them in a complementary manner. In addition, the data provided by CANCERTOOL represents a snapshot of gene expression levels in cancer specimens from patient cohorts recruited at different institutions and countries. We believe that the strength of CANCERTOOL resides in the capacity to browse through different patient cohorts for a given cancer type and seek for consistent alterations across Datasets. However, the data provided by CANCERTOOL should be complemented by experimental confirmation employing alternative methodological and analytical strategies.

I wish that CANCERTOOL will continue to support curiosity-driven cancer research.

Arkaitz Carracedo

2022:

- All datasets were downloaded again from the original sources (indicated in the Datasets section) to incorporate any update or modification (Feb).

- Latest normalization guidelines were applied (Feb):

  - For Datasets downloaded from GEO: normalization was performed by original authors, and no additional normalization process was applied upon download, their corresponding normalization is described in the corresponding GEO entry. Log2 has only been calculated for those datasets that required it after a thorough analysis of the data.
  - For Datasets downloaded from cBioportal (www.cbioportal.org): As described in the cBioPortal FAQ, for TCGA datasets "TCGA is processed and normalized using RSEM. Specifically, the RNASeq V2 data in cBioPortal corresponds to the rsem.genes.normalized_results file from TCGA", thus, no additional normalization process was applied upon download. For Metabric, it has been corroborated on the platform that the accessible data have already been normalized by the authors.

- The 2021 HGNC approved nomenclature was applied (Feb).

- The gene expression estimation for genes with more than one probe was calculated as the average signal of all available probes (Feb).

-We detected that the Metabric breast cancer dataset provided overall survival information mislabelled as disease-free survival and corrected the issue (Feb).

- Saioa Garcia Longarte was added to CANCERTOOL developing team (Feb).

-We verified the log transformation in mRNA expression data downloaded from *cbioportal* for Metabric and amended the data accordingly (Aug).