Balder raw data sources

Data source overview

This document describes the various raw data sources used to generate the methods and results within this repository. A number of publically available resources are utilized. Some resources allow for direct download for anyone with an internet connection. Other resources require users to register and agree to various terms of use. For this reason, data is not directly re-distributed with this code.

Biomarker actionability databases

CIViC

Clinical Interpretation of Variants in Cancer (CIViC) is an oper-source, community driven variant annotation platform. CIViC creates data releases on a fixed schedule. This study utilizes data from the "CIViC-01-Dec-2021" release. The data can be downloaded here:

https://civicdb.org/releases/main

Molecular Oncology Almanac

The Molecular Oncology Almanac (MOA) is a knowledge base for multimodal genomic biomarker actionability published in 2021. That data utilized in this study can be downloaded from the supplementary table found in the source paper. Please see Supplementary Tables Supplementary Tables 1–7, tab, "TableS2-MOAlmanac" from the article here:

https://www.nature.com/articles/s43018-021-00243-3#Sec20

Cancer Biomarker Database

The resourcces was last updated on 2022/10/17. A copy of the source table, "cgi_biomarkers_latest.tsv" was obtained on 12/19/2023 from the website here:

https://www.cancergenomeinterpreter.org/biomarkers

The following restrictions apply to these data: "This database is licensed under a Creative Commons Public Domain Dedication (CC0 1.0 Universal). When referring to this database, please cite: Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations; doi: https://doi.org/10.1101/140475."

Genomic data resources

MSK-IMPACT

Mutational profiles of over 10,000 patients were described in the study Zehir et. al, 2017 (DOI: 10.1038/nm.4333). Data from the 2017 study, "MSK-IMPACT Clinical Sequencing Cohort (MSK, Nat Med 2017", were downloaded from cBioPortal here:

https://www.cbioportal.org/

TCGA MC3

The MC3 variant data set is a compiled from TCGA data using multiple genomic pipelines. The data can be downloaded here:

https://gdc.cancer.gov/about-data/publications/mc3-2017

Barcode information for samples can be found here:

https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

Tissue source sites for each sample can be found here:

https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes

https://stephenturner.github.io/tcga-codes/

AACR GENIE

The data were downloaded as "Release v14.1 Public" from the Synapse website here:

https://www.synapse.org/#!Synapse:syn7222066/files/

Hartwig

Whole-genomie sequencing data for over 2,500 patients is described in the paper by Priestley, et. al, 2019 (https://doi.org/10.1038/s41586-019-1689-y). Access to these data requires an application with the Hartwig Medical Foundation:

https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/

ICGC TCGA WGS 2020

Whole genome data from over 2,600 patients from 38 tumor types. Publication of the data in 2020 is described by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (https://doi.org/10.1038/s41586-020-1969-6). Data were downloaded from the cBioPortal website, as “Pan-cancer analysis of whole genomes (ICGC/TCGA, Nature 2020)”:

https://www.cbioportal.org/