


# GenomeCruzer, a 3D Interactive Environment for Genomic Data Visualization and Analysis

Cassisa Anna<sup>1,2</sup><sup>a</sup>, Jamal Elhasnaoui<sup>1,2</sup>, Uliveto Chiara<sup>1,2</sup>, Riccardo Corsi<sup>3</sup>, Elena Grassi<sup>1,2</sup>, Dalibor Stuchlík<sup>4</sup>, Livio Trusolino<sup>1,2</sup>, Aleš Křenek<sup>4</sup>, Luca Vezzadini<sup>3</sup>, Andrea Bertotti<sup>1,2</sup>, Claudio Isella<sup>1,2</sup> and Enzo Medico<sup>1,2</sup>

<sup>1</sup>University of Torino, Department of Oncology, Candiolo, Torino, 10060, Italy

<sup>2</sup>Candiolo Cancer Institute, Fondazione Piemontese per l'Oncologia-Istituto di Ricovero e Cura a Carattere Scientifico (FPO-IRCCS), Candiolo, 10060, Italy

<sup>3</sup>Kairos3D, via Agostino da Montefeltro 2, 10134, Turin, Italy

<sup>4</sup>Institute of Computer Science, Masaryk University, Šumavská 15, 60200, Brno, Czech Republic

**Keywords:** GenomeCruzer, Genomic Landscape, View Mode, Graphical Metaphors, Multidimensional Omics, Colorectal Cancer, Breast Cancer.

**Abstract:** The development of high-throughput sequencing technologies has generated vast amounts of multi-layered molecular data from human tumours, but effectively visualizing and analysing these complex datasets remains a significant challenge for researchers. We introduce GenomeCruzer, a software designed to enable real-time, interactive visualization and analysis of large, multi-layer genomic and clinical data. GenomeCruzer uses graphical metaphors to represent continuous variables like gene expression, DNA methylation, and copy number alterations (CNA) through 3D objects with varying colour, size, and transparency, while discrete variables are represented by highlighting or blinking. We applied GenomeCruzer to DNA methylation and DNA/RNA sequencing data from colorectal cancer (CRC) samples from The Cancer Genome Atlas (TCGA) and CRC Patient-Derived Xenografts (PDXs). The software successfully generated 3D landscapes, allowing intuitive exploration of associations between omic profiles and clinical features. GenomeCruzer demonstrates its utility in highlighting subgroup differences, selecting representative cases, annotating samples, and identifying relationships between sample groups and gene signatures. Its intuitive interface and ability to visualize complex data make it a valuable tool for biomedical research.

## 1 INTRODUCTION

The implementation of Next Generation Sequencing (NGS) technologies in cancer research and clinical practice has advanced significantly over recent decades, leading to the generation of increasingly complex omics data (Dunn Jr et al., 2017). While these data are rich in information, they remain only partially explored due to their inherent complexity, posing challenges for non-specialist users, such as clinicians and biologists, who might otherwise utilize them for treatment decision-making or hypothesis generation (He et al., 2017). This complexity arises not only from the sheer volume of data but also from the intricate interactions across various layers of cell

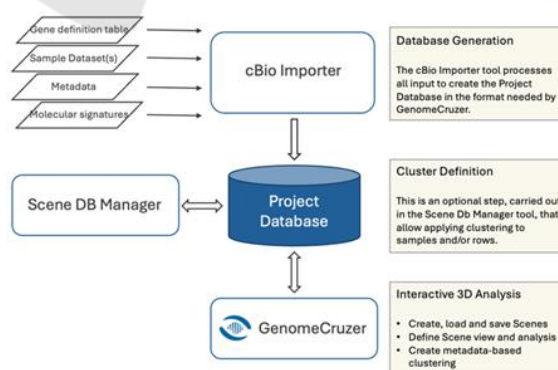



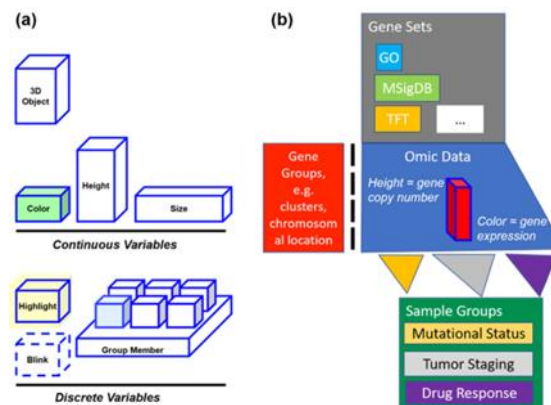
Figure 1: GenomeCruzer workflow.

 <https://orcid.org/0009-0007-7919-4598>

biology. Consequently, novel integrative approaches are needed to effectively link genes with samples, phenotypes, and other gene groups (Subramanian et al., 2020). The visualization of genomic data has emerged as a key discipline to enhance understanding and facilitate biological interpretation (Gao et al., 2013). A range of dedicated tools has been developed, employing diverse visualization methods from simple bar plots to sophisticated diagrams (Jia, 2011). However, each of these tools addresses specific challenges, which limits their ability to comprehensively characterize and interpret the data under investigation. For example, while widely utilized in various research fields, phylogenetic information (Partridge, 2018; West et al., 2006), clustered heatmaps (Eisen et al., 1998), and hierarchical or fuzzy clustering methods (Fu & Medico, 2007; McConnell et al., 2002) lack the capacity to integrate or simultaneously analyse multiple data types. Recently, more advanced tools such as cBioPortal (Gao et al., 2013; West et al., 2006), Complex Heatmaps (Gu et al., 2016), Integrative Genome Viewer (IGV) (Thorvaldsdóttir et al., 2013), the UCSC Xena platform (Goldman et al., 2020), Circos and its derivatives (Cui et al., 2020; Krzywinski et al., 2009), have been developed to enable enhanced visualization of omics data and to derive insights into specific cases (West et al., 2006). However, these methods often either focus on visualizing limited relationships across many samples, thereby reducing feature complexity, or provide detailed analyses of only one or a few samples at a time.

In this work, we investigate the potential of capturing multidimensional data, introducing GenomeCruzer, a software platform that provides a multidimensional 3D environment for the visualization of large-scale datasets using graphical metaphors. To clarify the workflow of the proposed method, we provide a flowchart (Figure 1), which illustrates the data flow within the GenomeCruzer software. The flowchart enables a clearer understanding of the integration and visualization process, offering readers a comprehensive overview of the steps involved in using the platform effectively. Conceptually, multiple variables associated with a single entity can be encoded as specific properties of a 3D object. For instance, continuous variables such as gene expression and copy number levels can be visualized simultaneously as colour and height, respectively, on a 3D parallelepiped representing the gene (Figure 2a). To derive meaningful insights from complex multidimensional data, metaphor encoding should be optimized to enable the effective

representation of objects, their variables, and associations within the 3D environment, while accommodating large cohorts of genes and samples without sacrificing resolution.



(a) Genomic variables are encoded as attributes of 3D objects: height, colour, and size for continuous variables (e.g. gene expression, copy number alteration, methylation levels), and highlights, blinks or group membership for discrete variables (e.g. gene mutation, sample/gene groups).  
 (b) Representation of the View Mode of multidimensional data in GenomeCruzer. The 3D space is divided into a floor and a wall. On the floor are represented the multi-omics data, with the values of their variables encoded as described in (a). Omics data can be hierarchically clustered into gene groups on rows and sample groups on columns.

Figure 2: GenomeCruzer concept for the visualisation of multidimensional genomic data with graphical metaphors.

This can be achieved by organizing 3D objects representing different entities in distinct spaces, such as the floor and walls of a virtual room (Figure 2b). GenomeCruzer allows users to perform intuitive, interactive analyses of diverse genomic data types and delivers real-time results as visual 3D landscapes. These features make GenomeCruzer a user-friendly interface for individuals with limited bioinformatics expertise, enabling them to explore and interpret extensive genomic datasets while gaining valuable insights for basic, translational, and clinical research.

## 2 RESULTS

### 2.1 Implementation of 3D Graphical Metaphors

GenomeCruzer is a software tool designed for the simultaneous visualization and analysis of multiple genomic datasets within a 3D interactive environment. By employing diverse graphical

metaphors, various genomic data types are represented as 3D objects. Continuous variables are visualized through attributes such as colour, height, and size, while discrete variables are represented through grouping, blinking, or highlighting.

This innovative visualization approach facilitates the identification of significant correspondences between genomic data types via a user-friendly interface (“View mode,” Figure 2b). Here, we present two demo datasets that allow users to explore the GenomeCruzer’s key features: (i) 570 colorectal cancer (CRC) samples from the TCGA-CRC dataset (Supplementary Table 1) (Muzny et al., 2012) and (ii) an in-house dataset comprising 53 CRC patient-derived xenograft (PDX) samples from the EuroPDX project (Supplementary Table 2) (Dudová et al., 2022). Users can reproduce results by connecting to the databases stored in the “GenomeCruzer\_DBs” subfolder within the GenomeCruzer Public Data Archive folder and selecting the corresponding scenes. Detailed instructions for recreating these databases are provided in the supplementary material, located in the “Source\_data” subfolder.

## 2.2 The View Mode: From Complexity to Clarity

GenomeCruzer’s 3D environment enables the simultaneous, interactive visualization of multiple layers of genomic data across large cohorts of samples. This feature allows users to intuitively explore potential correlations between different genomic layers, such as gene copy number alterations (CNA) and expression levels. By providing a comprehensive genomic landscape, the software allows users to investigate the relationships between various measurements of each gene within and across samples (see also Supplementary Video 1). To illustrate the innovative features of the View Mode, two representative use cases are presented:

1. A genome-wide landscape showcasing gene expression and CNA profiles for 570 colorectal cancer (CRC) samples obtained from The Cancer Genome Atlas (TCGA) (<https://www.cancer.gov/tcga>), organized by microsatellite instability (MSI) status, a well-established molecular marker in CRC (Pawlik et al., 2004).
2. A kinome-wide landscape displaying gene expression and CNA profiles for 53 CRC patient-derived xenograft (PDX) models, categorized by mutations in the RAS pathway and their in vivo

response to epidermal growth factor receptor (EGFR) blockade.

These examples highlight the versatility of GenomeCruzer’s View Mode in enabling researchers to uncover meaningful relationships and patterns across diverse genomic datasets, facilitating deeper insights into complex biological phenomena.

### 2.2.1 View Mode, Use Case 1: GenomeCruzer Highlights Chromosomal Domains with Recurrent and Concordant Differences in Gene Expression and CNA Between MSI and MSS CRCs

This use case demonstrates GenomeCruzer’s ability to highlight chromosomal domains with recurrent and concordant differences in gene expression and copy number alterations (CNA) between microsatellite instability (MSI) and microsatellite-stable (MSS) CRCs. Users can reproduce this analysis by connecting GenomeCruzer to the ‘TCGA\_CRC\_gcp\_cna.db’ database and selecting the scene titled ‘TCGA\_CRC:MSI\_MSS\_GenLandscape\_Custom\_genesets’. In this scene, the View Mode floor (Figure 3a) represents 570 TCGA CRC samples as columns, partitioned into four subgroups based on MSI status: MSI-high (MSI-H, n=77), MSI-low (MSI-L, n=91), microsatellite stable (MSS, n=397), and unclassified (NC, n=5) (Supplementary Table 1a). Genes are arranged as rows and hierarchically clustered by their genomic locations at five levels of increasing resolution (Supplementary Figure 1): chromosome, chromosomal arm, chromosomal band, and chromosomal sub-band, where each gene is represented as a single object. Gene expression and CNA profiles are displayed using graphical metaphors. Gene expression is represented by object colour: green for underexpression and red for overexpression (scaled as a log<sub>2</sub> ratio relative to the mean). CNA profiles are visualized as object heights: positive heights represent copy number gains, while negative heights correspond to copy number losses. Users can adjust the “Scale Factor” to modulate object height, ranging from 1 (no height) to 100 (maximum height) (Figure 3b). Further customization includes the “Absolute Value” option, which converts all heights to positive values directed above the floor (Figure 3b). When disabled, negative CNA values appear as objects extending below the floor (Figure 3c). The “Flip” option allows users to invert the orientation of positive and negative heights (Figure

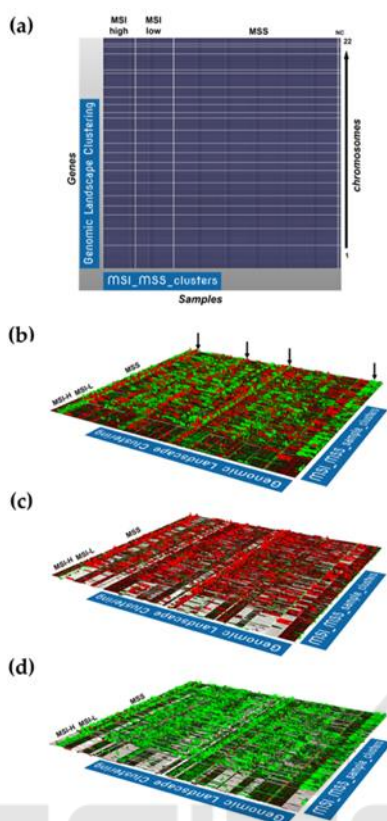


Figure 3: GenomeCruzer View Mode: the floor.

3d). At any resolution, the color and height of each object reflect the average gene expression and CNA values of its constituent genes. Users can select their desired resolution using the “Row Cluster Level” settings, ranging from chromosomes (low resolution) to individual genes (high resolution). Similarly, sample clustering can be adjusted with the “Column Cluster Level” settings, from broad classifications to more granular subdivisions of the samples. This genomic landscape visualization is particularly effective for assessing alterations across chromosomal domains

while maintaining resolution at the sample level. Users can quickly identify chromosomal domains with concordant differential gene expression and CNA profiles between MSI-H and (MSI-L+MSS) samples (Figure 3b-d). For instance, the 3D object heights clearly indicate a lower CNA burden in the MSI-H subgroup compared to the (MSI-L+MSS) subgroups (Figure 3b-d). Recurrent chromosomal alterations, such as gains and losses in regions like chr20q, chr18q, chr13q, and chr1q, can be visualized without compromising sample-specific details. Additionally, this mode allows for the simultaneous evaluation of gene expression and CNA levels in these regions (Figure 3b).

### 2.2.2 View Mode, Use Case 2: Kinome Visualisation in GenomeCruzer Reveals Concomitant Amplifications and Over-Expressions of Therapeutic Target Kinases in PDX Models

This use case demonstrates the potential of GenomeCruzer to identify actionable therapeutic targets by simultaneously visualizing mRNA expression and copy number alteration (CNA) profiles across large cohorts of patient-derived xenograft (PDX) models.

Users can replicate this analysis by connecting GenomeCruzer to the ‘PDX\_CRC\_gep\_cna.db’ database and selecting the ‘RAS\_Kinome\_families\_Custom\_genesets’ scene. The clinical translation of genomic assays and the development of targeted therapies for cancer treatment rely heavily on identifying actionable target genes. This requires an in-depth investigation into molecular alterations that could be under selective pressure at both genomic and transcriptional levels, potentially leading to the discovery of novel druggable targets (Jeon et al., 2014; Tran & Pham, 2021).

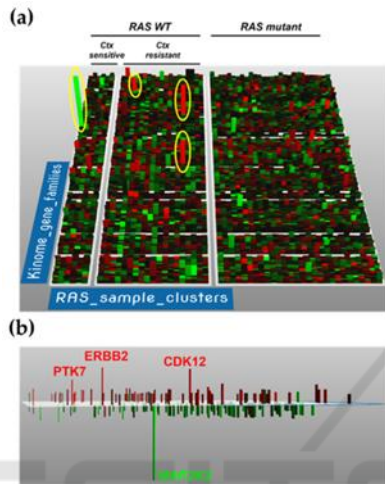
This specific scene illustrates how GenomeCruzer facilitates the identification of such targets by enabling users to explore their mRNA expression and CNA profiles. To highlight this functionality, we focused on kinase gene families, which are particularly promising candidates for pharmacological inhibition in cancer due to their central roles in cell signaling pathways. These families often exhibit sparse, outlier genomic alterations within specific subgroups of samples (Medico et al., 2015).

The dataset used in this analysis consists of 53 PDXs annotated for their response to cetuximab, an anti-EGFR monoclonal antibody and a key therapeutic option for CRC (Bertotti et al., 2011).



Given that cetuximab efficacy is compromised by mutations in downstream effectors of EGFR, only RAS/RAF wildtype cases are eligible for cetuximab treatment. Accordingly, the 53 PDXs were annotated for hotspot mutations in KRAS, NRAS, and BRAF, which are downstream kinases of EGFR.

The PDXs were hierarchically classified into three groups: RAS wildtype cetuximab-sensitive (n=6), RAS wildtype cetuximab-resistant (n=19), and RAS mutant (n=28) (Supplementary Table 2a).



(a) Front view of the kinome floor, with rows representing kinases hierarchically clustered by family and columns depicting 53 PDX samples grouped by cetuximab sensitivity and RAS mutational status. Each cell indicates gene expression (color) and CNA value (height, absolute value). Yellow ovals highlight specific kinases with notable expression and CNA values.

(b) Side view of the kinome floor, showing positive CNA values (upper side) and negative CNA values (lower side). The “hide subrange” function excludes genes with minor alterations, emphasizing kinases with the most substantial changes.

Figure 4: GenomeCruzer View Mode visualizations of the kinome floor, showcasing “outlier” kinase expression and CNA in CRC PDXs.

Kinase-encoding genes were similarly clustered by family, following the classification scheme provided by Kinhub (kinhub, <http://www.kinhub.org/index.html>, accessed 15/04/2022). This classification includes ten families: (i) the Tyrosine Kinase (TK) and (ii) their Like (TKL) groups, (iii) the Serine/Threonine-protein kinases (ACG) group, (iv) the proline-directed Serine/Threonine (CMGC) kinase group, (v) the mitogen-activated and Serine/Threonine protein kinases (STE) group, (vi) the Ca<sup>2+</sup> calmodulin-dependent kinases (CAMK) group, (vii) the Casein Kinase 1 (CK1) group, (viii) the Receptor Guanylate Cyclase kinase (RCG) group, in addition to two other

orphan groups that were annotated as atypical or other. GenomeCruzer's visualizations effectively highlight genomic alterations relevant to therapeutic strategies. For example, ERBB2 (HER2) emerged as a highly expressed outlier in the RAS wildtype, cetuximab-resistant subgroup. This observation aligns with prior studies identifying ERBB2 as a mechanism of resistance to cetuximab and a promising therapeutic target currently under evaluation in phase II clinical trials for CRC treatment (Sartore-Bianchi et al., 2020).

Among the RAS wildtype cetuximab-resistant samples, the CRC0112 sample, documented in this study (Supplementary Table 2a), had been previously treated successfully with pertuzumab and lapatinib (Bertotti et al., 2011). Other notable outliers included LCK, a member of the SRC family implicated in receptor tyrosine kinase signaling and a target of Dasatinib (Lombardo et al., 2004), and PTK7, an atypical kinase whose alterations are associated with poor CRC prognosis (Tian et al., 2016).

In RAS mutant samples (Supplementary Table 2a), FLT1 and FGFR1 were identified as outlier receptor tyrosine kinases associated with tumor progression across multiple cancer types (Bae et al., 2019; Minev, 2011; Miyake et al., 2016; Slattery et al., 2014). These kinases represent promising therapeutic targets for patients not responding optimally to anti-EGFR therapies.

Additional targets include DYRK4, a less characterized kinase whose outlier overexpression and CNA levels suggest potential utility in combination treatments with cetuximab. Another example is MAP2K3, which exhibited substantial gene depletion and significantly reduced RNA expression levels in the RAS wildtype cetuximab-sensitive subgroup. Located on chromosome 17, MAP2K3 has been implicated in stage II colon cancer as a candidate driver of focal chromosomal aberrations (Brosens et al., 2010). Notably, MAP2K3 functions as a tumor suppressor in some cancers, such as breast cancer, but exhibits oncogenic roles in CRC (Piastra et al., 2022).

GenomeCruzer's “hide subrange” feature further enhances data interpretation by allowing users to filter out genes with minor alterations, highlighting kinases with the most pronounced changes in CNA and expression. For instance, this functionality pinpointed ERBB2, PTK7, and MAP2K3 as candidates with significant alterations, supporting their consideration as potential therapeutic targets.

Collectively, these results underscore GenomeCruzer's effectiveness in identifying putative therapeutic targets by integrating and visualizing

complex genomic datasets. Additionally, researchers can expand the analysis to other gene groups, such as transcription factors, splicing factors, or microRNAs, demonstrating the platform's versatility. Supplementary Video 2 provides a comprehensive demonstration of this use case.

### 3 DISCUSSION

Over the last two decades, large-scale application of NGS has generated vast amounts of genomic data, enabling patient stratification and diagnostic advancements (Das et al., 2020; Duan et al., 2021; Zhu et al., 2016), opening a new horizon in clinical diagnosis (Cantini et al., 2021). While NGS-based assays are increasingly used in cancer diagnosis and treatment (Mordente et al., 2015), integrating multiple omics with clinical data remains a major challenge (Gomez-Cabrero et al., 2014). Multi-omics research is expected to uncover interactions among molecular entities and improve disease outcome predictions (Berger & Mardis, 2018; Subramanian et al., 2020). However, integrating these datasets often requires complex data mining and machine learning, making tools accessible only to bioinformatics experts. GenomeCruzer addresses these challenges by providing an intuitive, 3D environment for multidimensional omics data integration and real-time analysis. It enables users to explore large-scale datasets, such as those from the EuroPDX Consortium (Dudová et al., 2022), and public repositories like TCGA and cBioPortal (Gao et al., 2013; West et al., 2006), without requiring bioinformatics skills. Of note, cBioPortal offers the opportunity to explore very large sample cohorts, but only for a limited number of genes. Beyond a certain number of genes/features, the 2d nature of the display does not provide sufficient resolution. Conversely, a key feature of GenomeCruzer is its Genomic Landscape, which enables wide-scale exploration of genomic data, from the whole genome down to single genes. This approach surpasses common tools like Integrative Genome Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013), cBioPortal (Gao et al., 2013; West et al., 2006), Circos (Cui et al., 2020; Krzywinski et al., 2009) and Xena (Goldman et al., 2020), by allowing the concurrent visualization of two omics layers. Circos and Complex Heatmaps, on the other hand, are highly effective for identifying patterns and relationships within circular or grid-based visualizations, but they lack the interactive and exploratory features of GenomeCruzer. The fixed nature of these representations makes them less suited

for dynamic hypothesis generation compared to GenomeCruzer's real-time data interaction capabilities. This is not possible with IGV or similar tools, which show only a portion of a genomic region, across a limited number of samples. This limit is not a choice, it is rather a consequence of the visualization environment. This feature facilitates the identification of genomic lesions, recurrent chromosomal alterations, and the relationship between different genomic data layers. It is also more effective than conventional tools like GISTIC (Mermel et al., 2011) in detecting differential gene expression/CNA/methylation across sample groups. By running GenomeCruzer on TCGA colon and breast cancer datasets, we identified known chromosomal alterations, demonstrating its power in complex cancer analyses. In summary, GenomeCruzer offers an innovative tool for multi-omics data analysis, simplifying complex cancer biology exploration without requiring specialized computational expertise. Its unique capabilities could enhance both research and clinical applications of genomic data.

### 4 CONCLUSIONS

*GenomeCruzer* provides a user-friendly, flexible, innovative, and powerful tool for 3D visualisation, integration, and interpretation of multi-dimensional genomic data. As the field of omics data continues to expand and more public omics datasets are available, GenomeCruzer will evolve as an integrated solution for deciphering underlying biological information. GenomeCruzer is currently distributed as shareware and can be freely used.

### 5 SUPPLEMENTARY INFORMATION

#### 5.1 Availability of Data and Material

All datasets analysed in this study including TCGA-CRC and PDXs are available in the TCGA (<https://www.cancer.gov/tcga>) and the EuroPDX (<https://www.europdx.eu/>) data repositories, respectively. The showcase Genomic databases, and the raw data to build these databases can be found at GenomeCruzer Public Data Archive. All the Supplementary Materials are available from <https://github.com/acassisa/BIOINFORMATICS2025>.

## 5.2 GenomeCruzer Software

For the version of GenomeCruzer available at the time of this publication, please write to genomecruzer@kairos3d.it or use the following link: <http://www.genomecruzer.com>.

Project name: GenomeCruzer

Project home page: <http://www.genomecruzer.com>

Operating system(s): Windows. GenomeCruzer is available now on Windows. On request, it can be made available also for Ubuntu and MacOS users.

Programming language: C/C++

Other requirements: None

Licence: Custom freeware licence

Any restrictions to use by non-academics: None

## 5.3 Preprocessing Instructions for cBioportal Data

This manual outlines the steps to generate a comprehensive database using RNA and CNA datasets from cBioPortal, a widely used platform for cancer genomics data. The process involves selecting, formatting, and integrating the necessary files to ensure compatibility with subsequent analysis.

### Data Retrieval

The process begins with collecting the required data from cBioPortal. After identifying the dataset of interest, download the following key files: 1. **data\_mrna\_seq.txt** – Contains mRNA sequencing data. 2. **data\_cna.txt** – Contains copy number alteration (CNA) data. 3. **data\_clinical\_sample.txt** – Contains clinical sample annotations. These files must be in the standard formats used by cBioPortal, as the subsequent steps rely on this consistency.

### Database Creation Workflow

The database generation process involves six main steps: 1. **Definition of the Sample Dataset** Select and curate the sample-level data, ensuring alignment with the scope of the analysis. 2. **Definition of the genomic annotation Database** Define the set of genes to be included, based on the study objectives or predefined panels. 3. **Definition of Molecular Data** Process the mRNA and CNA data to ensure they are ready for integration. This may involve normalization or reformatting as needed. 4. **Database Generation** Integrate the sample, gene, and molecular data into a unified database. 5. **Sample Cluster Annotation** Perform clustering analysis on the samples and annotate the resulting groups to identify meaningful patterns. 6. **Gene Cluster Annotation** Conduct clustering analysis on the genes and provide

annotations to highlight biological relevance or pathways of interest.

### Database File Format

1. **Definition of Samples' Dataset:** From the `data_clinical_sample.txt` file, the following columns are retained, and the samples list is created:

```
#Patient Identifier   Sample Identifier
#Identifier to uniquely specify a patient.   A unique sample identifier.
#STRING             STRING
#1                 1
PATIENT_ID        SAMPLE_ID
```

2. **Definition of the Genomic Annotation Database:** The genomic annotation database is structured as a tab delimited csv file using `UNIQUE_ID` as entry. The database contains a numeric `UNIQUE_ID` and an associated `GENE_ID`, together with the chromosome location of the genes and their start position, end position, `ARM_ID`, `BAND_ID` and `SUB BAND ID`. A default genic database of 19603 genes is used based on HG19.

3. **Definition of the Molecular Data. RNA matrix:** Duplicates removal is performed and the numerical entrez id should be the first column of the file. In the same folder, the meta-data for this dataset is created as follows: “`cancer_study_identifier`” is the name of the study, “`genetic_alteration_type`” is set to `mRNA_expression`, “`datatype`” is set to `continuous`, “`stable_id`” is set to `rna_seq_mrna_capture`, “`show_profile_in_analysis_tab`” is set to `true`, “`profile_name`” is the mRNA expression (ILMN-Linear), “`profile_description`” is set to `Expression levels (Log2, RNAseq)` and “`data_filename`” to `data_mrna.tsv`.

**CNA matrix:** The matrix is generated starting from segmentation data using the `code_cna.R` script. This code takes as an input the genic dataset and the segmentation data and after extracting unique genes and samples, CNA values are collected for the corresponding segments. Whereas multiple segment are attributed to a gene in a sample, the CNA value is calculated as the average expression of such segment normalized on the length of the segment included in the genomic region. The meta-data is created

for this dataset as well, as follows: “cancer\_study\_identifier” is the name of the study, “genetic\_alteration\_type” is set to COPY\_NUMBER\_ALTERATION, “datatype” is LOG2-VALUE, “stable\_id” is set to log<sub>2</sub>CNA, “show\_profile\_in\_analysis\_tab” is set to true, “profile\_name” is Log<sub>2</sub> copy-number values, “profile\_description” is the Copy-number values for each gene (from shallow seq) and “data\_filename” is cna\_matrix.tsv. methylation matrix The matrix is generated starting from **beta values** data using the *code\_methylation.R* script. This code takes as an input the RNA\_matrix dataset and the meth\_data, and after extracting unique genes and samples, b values values are attributed to genes selecting the locus with the highest absolute correlation for the gene expression values. collected for the segments, even in the case. The meta-data is created for this dataset as well, as follows: “cancer\_study\_identifier” is the name of the study, “genetic\_alteration\_type” is set to Methylation, “datatype” is LOG2RATIO, “stable\_id” is set to “methylation\_hm450”, “show\_profile\_in\_analysis\_tab” is set to true, “profile\_name” is Log<sub>2</sub> copy-number values, “profile description” is the beta value for each gene and “meth\_filename” is cna\_matrix.tsv.

4. Database Generation: The database is generated using a json file, containing the instructions and the paths to the files needed. A template is available for the creation of this file. Through the use of a prompt, the database is generated. Once in the folder where the database should be created, the following instructions are given to cBioImporter.exe program:  

```
...\\Genome Cruzer\\1.9.0\\bin>cBioImporter.exe -j nuovo_db\\code\\database_cBioImport.json --out nuovo_db\\Outputs\\database.db
```
5. Sample Cluster Annotation: After the database has been created, we could provide a sample clustering, given as column clustering onto DbManager.exe program. This clustering takes the form of a TAB delimited text file, containing all the samples in the database (samples do not present in the database should not be listed). An

example for the heading of this file is shown below:

```
COLUMNS 2
UNIQUE_ID CLUS_ID_1 CLUS_ID_2
```

6. Gene Cluster Annotation: Gene clustering is provided as row clustering onto DbManager.exe program. Some pre-defined files could be found in the clusterings folder, as an example.

## REFERENCES

- Bae, J. M., Wen, X., Kim, T.-S., Kwak, Y., Cho, N.-Y., Lee, H. S., & Kang, G. H. (2019). Fibroblast Growth Factor Receptor 1 (FGFR1) Amplification Detected by Droplet Digital Polymerase Chain Reaction (ddPCR) Is a Prognostic Factor in Colorectal Cancers. *Cancer Res Treat*, 52(1), 74–84. <https://doi.org/10.4143/crt.2019.062>
- Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, 15(6), 353–365. <https://doi.org/10.1038/s41571-018-0002-6>
- Bertotti, A., Migliardi, G., Galimi, F., Sassi, F., Torti, D., Isella, C., Corà, D., di Nicolantonio, F., Buscarino, M., Petti, C., Ribero, D., Russolillo, N., Muratore, A., Massucco, P., Pisacane, A., Molinaro, L., Valtorta, E., Sartore-Bianchi, A., Risio, M., ... Trusolino, L. (2011). A molecularly annotated platform of patient-derived xenografts (“xenopatiens”) identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discovery*, 1(6), 508–523. <https://doi.org/10.1158/2159-8290.CD-11-0109>
- Brosens, R. P. M., Haan, J. C., Carvalho, B., Rustenburg, F., Grabsch, H., Quirke, P., Engel, A. F., Cuesta, M. A., Maughan, N., Flens, M., Meijer, G. A., & Ylstra, B. (2010). Candidate driver genes in focal chromosomal aberrations of stage II colon cancer. *The Journal of Pathology*, 221(4), 411–424. <https://doi.org/https://doi.org/10.1002/path.2724>
- Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1), 124. <https://doi.org/10.1038/s41467-020-20430-7>
- Cui, Y., Cui, Z., Xu, J., Hao, D., Shi, J., Wang, D., Xiao, H., Duan, X., Chen, R., & Li, W. (2020). NG-Circos: next-generation Circos for data visualization and interpretation. *NAR Genomics and Bioinformatics*, 2(3), lqaa069. <https://doi.org/10.1093/nargab/lqaa069>
- Das, T., Andrieux, G., Ahmed, M., & Chakraborty, S. (2020). Integration of Online Omics-Data Resources for Cancer Research. *Frontiers in Genetics*, 11.



- <https://www.frontiersin.org/journals/genetics/article/s/10.3389/fgene.2020.578345>
- Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., Zhang, C., & Jia, S. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Computational Biology*, 17(8), e1009224. <https://doi.org/10.1371/journal.pcbi.1009224>
- Dudová, Z., Conte, N., Mason, J., Stuchlík, D., Peša, R., Halmagyi, C., Perova, Z., Mosaku, A., Thorne, R., Follette, A., Pivarč, L., Šašinka, R., Usman, M., Neuhauser, S., Begley, D. A., Krupke, D. M., Frassà, M., Fiori, A., Corsi, R., ... Křenek, A. (2022). The EurOPDX Data Portal: an open platform for patient-derived cancer xenograft data sharing and visualization. *BMC Genomics*, 23(1), 156. <https://doi.org/10.1186/s12864-022-08367-1>
- Dunn Jr, W., Burgun, A., Krebs, M.-O., & Rance, B. (2017). Exploring and visualizing multidimensional data in translational research platforms. *Briefings in Bioinformatics*, 18(6), 1044–1056. <https://doi.org/10.1093/bib/bbw080>
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863–14868. <https://doi.org/10.1073/pnas.95.25.14863>
- Fu, L., & Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, 8(1), 1–15. <https://doi.org/10.1186/1471-2105-8-3>
- Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269), e245. <https://doi.org/10.1126/scisignal.2004088>
- Goldman, M. J., Craft, B., Hastie, M., Repčeka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., & Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38(6), 675–678. <https://doi.org/10.1038/s41587-020-0546-8>
- Gomez-Cabrero, D., Abugensaisa, I., Maier, D., Teschendorff, A., Merckenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., & Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2), I1. <https://doi.org/10.1186/1752-0509-8-S2-I1>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847–2849. <https://doi.org/10.1093/bioinformatics/btw313>
- He, K. Y., Ge, D., & He, M. M. (2017). Big Data Analytics for Genomic Medicine. *International Journal of Molecular Sciences*, 18(2), 24853. <https://doi.org/10.3390/ijms18020412>
- Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., Moffat, J., & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, 6(7), 57. <https://doi.org/10.1186/s13073-014-0057-7>
- Jia, M. (2011). *Visualizing Biological Data in Google Earth*. Iowa State University. <https://books.google.it/books?id=qOKJAQAACAAJ>
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645. <http://genome.cshlp.org/content/19/9/1639.abstract>
- Lombardo, L. J., Lee, F. Y., Chen, P., Norris, D., Barrish, J. C., Behnia, K., Castaneda, S., Cornelius, L. A. M., Das, J., Doweyko, A. M., Fairchild, C., Hunt, J. T., Inigo, I., Johnston, K., Kamath, A., Kan, D., Klei, H., Marathe, P., Pang, S., ... Borzilleri, R. M. (2004). Discovery of N-(2-Chloro-6-methyl-phenyl)-2-(6-(4-(2-hydroxyethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), a Dual Src/Abl Kinase Inhibitor with Potent Antitumor Activity in Preclinical Assays. *Journal of Medicinal Chemistry*, 47(27), 6658–6661. <https://doi.org/10.1021/jm049486a>
- McConnell, P., Johnson, K., & Lin, S. (2002). Applications of Tree-Maps to hierarchical biological data. *Bioinformatics*, 18(9), 1278–1279. <https://doi.org/10.1093/bioinformatics/18.9.1278>
- Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B., Veronese, S., Siena, S., Sartore-Bianchi, A., Beccuti, M., Mottolese, M., Linnebacher, M., Cordero, F., Di Nicolantonio, F., & Bardelli, A. (2015). The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nature Communications*, 6. <https://doi.org/10.1038/ncomms8002>
- Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4), 1–14. <https://doi.org/10.1186/gb-2011-12-4-r41>
- Minev, B. (2011). *Cancer Management in Man: Chemotherapy, Biological Therapy, Hyperthermia and Supporting Measures*. <https://doi.org/10.1007/978-90-481-9704-0>
- Miyake, T., Kumasawa, K., Sato, N., Takiuchi, T., Nakamura, H., & Kimura, T. (2016). Soluble VEGF receptor 1 (sFLT1) induces non-apoptotic death in ovarian and colorectal cancer cells. *Scientific Reports*, 6(1), 24853. <https://doi.org/10.1038/srep24853>

- Mordente, A., Meucci, E., Martorana, G. E., & Silvestrini, A. (2015). Cancer Biomarkers Discovery and Validation: State of the Art, Problems and Future Perspectives. In R. Scatena (Ed.), *Advances in Cancer Biomarkers: From biochemistry to clinic for a critical revision* (pp. 9–26). Springer Netherlands. [https://doi.org/10.1007/978-94-017-7215-0\\_2](https://doi.org/10.1007/978-94-017-7215-0_2)
- Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y. Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., ... Thomson, E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. <https://doi.org/10.1038/nature11252>
- Partridge, D. (2018). Darwin's two theories, 1844 and 1859. *Journal of the History of Biology*, *51*(3), 563–592. <https://doi.org/10.1007/s10739-018-9509-z>
- Pawlik, T. M., Raut, C. P., & Rodriguez-Bigas, M. A. (2004). Colorectal Carcinogenesis: MSI-H Versus MSI-L. *Disease Markers*, *20*(4–5), 368680. <https://doi.org/https://doi.org/10.1155/2004/368680>
- Piastra, V., Pranteda, A., & Bossi, G. (2022). Dissection of the MKK3 Functions in Human Cancer: A Double-Edged Sword? *Cancers*, *14*, 483. <https://doi.org/10.3390/cancers14030483>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Sartore-Bianchi, A., Lonardi, S., Martino, C., Fenocchio, E., Tosi, F., Ghezzi, S., Leone, F., Bergamo, F., Zagonel, V., Ciardiello, F., Ardizzoni, A., Amatu, A., Bencardino, K., Valtorta, E., Grassi, E., Torri, V., Bonoldi, E., Sapino, A., Vanzulli, A., ... Siena, S. (2020). Pertuzumab and trastuzumab emtansine in patients with HER2-amplified metastatic colorectal cancer: the phase II HERACLES-B trial. *ESMO Open*, *5*(5). <https://doi.org/10.1136/esmoopen-2020-000911>
- Slattery, M. L., Lundgreen, A., & Wolff, R. K. (2014). VEGFA, FLT1, KDR and colorectal cancer: Assessment of disease risk, tumor molecular phenotype, and survival. *Molecular Carcinogenesis*, *53*(S1), E140–E150. <https://doi.org/https://doi.org/10.1002/mc.22058>
- Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*, 1177932219899051. <https://doi.org/10.1177/1177932219899051>
- Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. <https://doi.org/10.1093/bib/bbs017>
- Tian, X.-Y., Yan, L., Zhang, D., Guan, X., Dong, B., Zhao, M., & Hao, C. (2016). PTK7 overexpression in colorectal tumors: Clinicopathological correlation and prognosis relevance. *Oncology Reports*, *36*, 1829–1836. <https://api.semanticscholar.org/CorpusID:17035375>
- Tran, T.-D., & Pham, D.-T. (2021). Identification of anticancer drug target genes using an outside competitive dynamics model on cancer signaling networks. *Scientific Reports*, *11*(1), 14095. <https://doi.org/10.1038/s41598-021-93336-z>
- West, M., Ginsburg, G. S., Huang, A. T., & Nevins, J. R. (2006). Embracing the complexity of genomic data for personalized medicine. *Genome Research*, *16*(5), 559–566. <http://genome.cshlp.org/content/16/5/559.abstract>
- Zhu, R., Zhao, Q., Zhao, H., & Ma, S. (2016). Integrating multidimensional omics data for cancer outcome. *Biostatistics*, *17*(4), 605–618. <https://doi.org/10.1093/biostatistics/kxw010>