

Introduction to cBioPortal

21/03/2024



Course material



Andrew Mason

BSc (Hons) PhD AFHEA



Lecturer in Cancer Informatics at The University of York

Run a small bioinformatics-focused research group within the Jack Birch Unit

Work on human urothelial cancers and retroviral cancers in birds

Bioinformatic lead for the bladder cancer group of the 100,000 genomes project

Elixir-UK Data Stewardship Training Fellow

Improve data management, not just analysis, in life sciences

Development of data management training resources



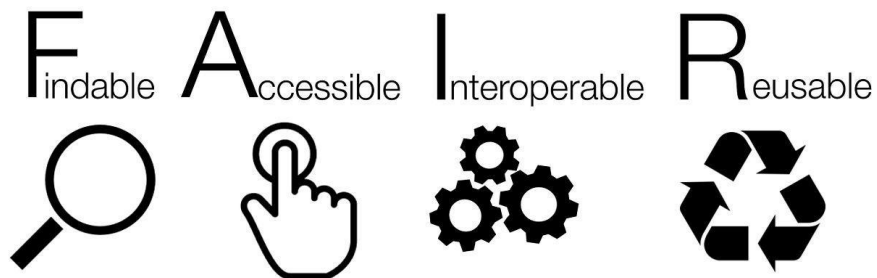
What is Elixir?

"ELIXIR coordinates and develops life science resources across Europe so that researchers can more easily find, analyse and share data, exchange expertise, and implement best practices."

Improve skills in data management

Improve quality of, and access to, informatics training

Develop and disseminate FAIR data principles in life sciences



Elixir-UK Data Stewardship Training Fellows



[Research Data Management bites](#)



[My videos introducing sequencing data](#)



Online training courses, cookbooks and carpentries



Local, in-person training

Introduction and Learning Objectives

Introduction to cBioPortal

Course material



Session structure

- 12.35 Introduction and Learning Objectives
- 12.45 cBioPortal website demonstration
- 12.55 Problem-solving tasks
- 13.30 Recap and Further Resources
- 13.35 Accessing and using underlying cBioPortal data
- 13.45 Rstudio demo for accessing cBioPortal data
- 13.55 Summary and final remarks
- 14.00 Close

Course material



Learning objectives

- 1 Recognise the applications and utility of cBioPortal for cancer research
- 2 Operate and explore the cBioPortal website to identify cancer data of interest
- 3 Complete two cancer biology problem-solving tasks using cBioPortal
- 4 Recognise the process for accessing and analysing cBioPortal data

Course material



What is cBioPortal?



Public website for exploratory analysis, visualisation and download of large cancer omics datasets, with clinical metadata

Data derived from large consortia, as well as highlighted studies



Data annotated by external reference databases

Depending on the dataset, includes mutations, CNA, gene expression, methylation data *etc.*

Course material



When using cBioPortal, cite the following papers, plus the reference papers for datasets used: [Cerami *et al.* 2012](#), [Gao *et al.* 2013](#)



Why use cBioPortal?



Access to the largest, publicly available cancer sequencing studies, all in one place

Explore broader relevance of laboratory/animal studies across cancer types

Hypothesis generation, including student projects

Data visualisation

Exploration of clinical data

Course material



When using cBioPortal, cite the following papers, plus the reference papers for datasets used: [Cerami *et al.* 2012](#), [Gao *et al.* 2013](#)



cBioPortal website demo

Introduction to cBioPortal

Course material



Problem-solving tasks

Your turn!

Course material



Google Form for questions




Recap and Further Resources

Introduction to cBioPortal

- 1 Recognise the applications and utility of cBioPortal for cancer research
- 2 Operate and explore the cBioPortal website to identify cancer data of interest
- 3 Complete two cancer biology problem-solving tasks using cBioPortal
- 4 Recognise the process for accessing and analysing cBioPortal data

Further training resources

 cBioPortal Docs

[cBioPortal](#) [Github](#) [Issues](#) [Google Group](#)

Filter

- Home
- About Us
- List of RFCs
- User Guide
 - New Users
 - Frequently Asked Questions
 - Overview of Resources**
 - Page Specific Resources
 - OQL
 - News
 - Genie News
 - Web API and API Clients
- Deployment >
- Development >

Overview

Overview of Resources

Tutorial Slides

These tutorial slides contain annotated screenshots to walk you through using the cBioPortal site.

1. Single Study Exploration [Google slides](#) | [PDF](#)
2. Single Study Query [Google slides](#) | [PDF](#)
3. Patient View [Google slides](#) | [PDF](#)
4. Virtual Studies [Google slides](#) | [PDF](#)
5. Onco Query Language (OQL) [Google slides](#) | [PDF](#)
6. Group Comparison [Google slides](#) | [PDF](#)
7. Pathways [Google slides](#) | [PDF](#)

Webinar Recordings

Recordings of live webinars from April & May 2020

1. Introduction to cBioPortal [youtube.com](#) | [bilibili.com](#) | [Download PDF](#) | [View slides](#)
2. Mutation Details & Patient View [youtube.com](#) | [bilibili.com](#) | [Download PDF](#) | [View slides](#)
3. Expression Data Analysis [youtube.com](#) | [bilibili.com](#) | [Download PDF](#) | [View slides](#)
4. Group Comparison [youtube.com](#) | [bilibili.com](#) | [Download PDF](#) | [View slides](#)
5. API & R Client [youtube.com](#) | [bilibili.com](#) | [Download PDF](#) | [View slides](#) | [Workshop code](#)

How-To Videos

Short videos that show how to perform specific analyses or how to use specific pages.

Direct link



cBioPortal FAQs



Accessing and using underlying cBioPortal data

Introduction to cBioPortal

4

Recognise the process for accessing and analysing cBioPortal data

Why bother?

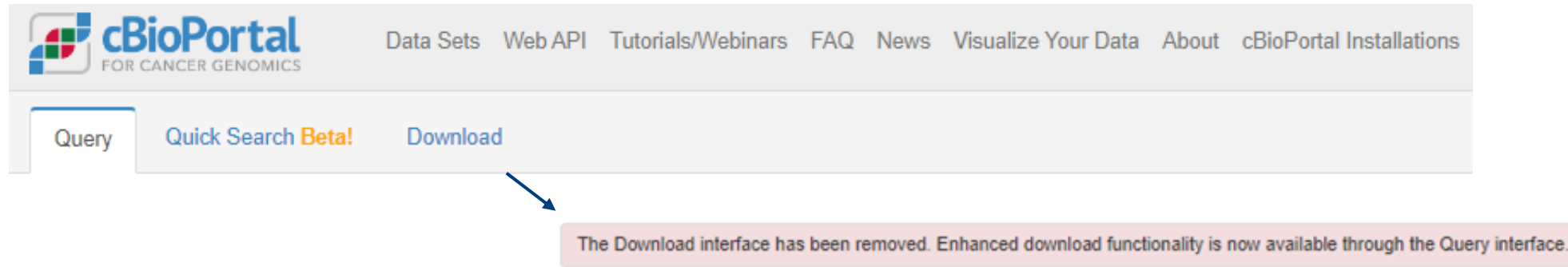
“under-the-hood” dataset has more information than displayed publicly

analyse lists of genes quickly

“improve” the plot quality

perform more advanced statistical testing (e.g. DEA, GSEA)

Downloading data

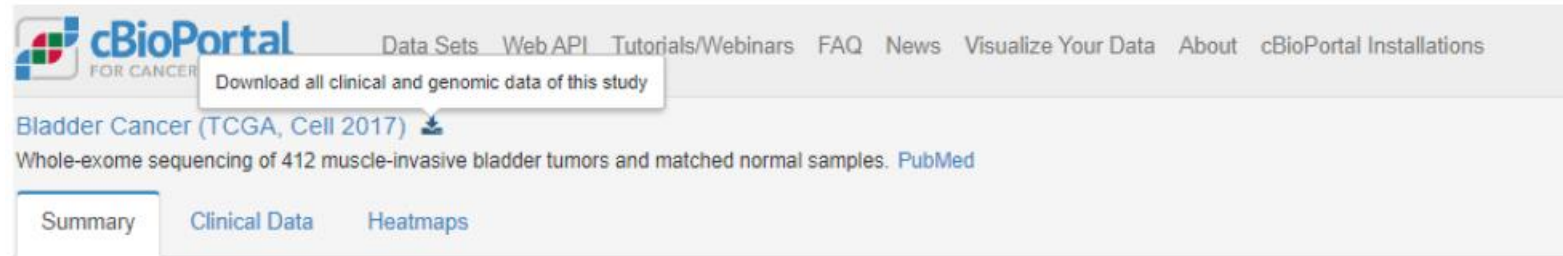


The screenshot shows the top navigation bar of the cBioPortal website. The logo for cBioPortal FOR CANCER GENOMICS is on the left. To its right are links for Data Sets, Web API, Tutorials/Webinars, FAQ, News, Visualize Your Data, About, and cBioPortal Installations. Below the navigation bar is a menu with three items: Query, Quick Search Beta!, and Download. A blue arrow points from the Download link to a light pink message box that reads: "The Download interface has been removed. Enhanced download functionality is now available through the Query interface."

Unhelpful starting point.

Downloading data

Explore your dataset first, and then download.



The screenshot shows the cBioPortal interface. At the top left is the cBioPortal logo with the tagline 'FOR CANCER'. To the right of the logo is a navigation menu with links: Data Sets, Web API, Tutorials/Webinars, FAQ, News, Visualize Your Data, About, and cBioPortal Installations. Below the logo is a button that says 'Download all clinical and genomic data of this study'. The main content area displays the dataset title 'Bladder Cancer (TCGA, Cell 2017)' with a download icon, followed by the description 'Whole-exome sequencing of 412 muscle-invasive bladder tumors and matched normal samples. PubMed'. At the bottom of the content area are three tabs: 'Summary' (which is selected), 'Clinical Data', and 'Heatmaps'.

Download will start and give a `.tar.gz` file

Downloading data

blca_tcga_pub_2017 ↕ ↻

Name	Date modified	Type	Size
case_lists	25/03/2022 19:07	File folder	
→ data_clinical_patient.txt	25/03/2022 19:15	TXT File	355 KB
→ data_clinical_sample.txt	25/03/2022 19:15	TXT File	103 KB
→ data_cna.txt	25/03/2022 19:15	TXT File	22,499 KB
data_linear_cna.txt	25/03/2022 19:15	TXT File	64,382 KB
data_methylation_hm450.txt	25/03/2022 19:15	TXT File	119,597 KB
→ data_mrna_seq_v2_rsem.txt	25/03/2022 19:15	TXT File	69,336 KB
data_mrna_seq_v2_rsem_zscores_ref_all_samples.txt	25/03/2022 19:15	TXT File	60,450 KB
data_mrna_seq_v2_rsem_zscores_ref_diploid_samples.txt	25/03/2022 19:15	TXT File	59,941 KB
→ data_mutations.txt	25/03/2022 19:15	TXT File	264,752 KB
data_mutsig.txt	25/03/2022 19:15	TXT File	2,083 KB
data_rppa.txt	25/03/2022 19:15	TXT File	643 KB
data_rppa_zscores.txt	25/03/2022 19:15	TXT File	569 KB
LICENSE	25/03/2022 19:07	File	1 KB
meta_clinical_patient.txt	25/03/2022 19:07	TXT File	1 KB
meta_clinical_sample.txt	25/03/2022 19:07	TXT File	1 KB
meta_cna.txt	25/03/2022 19:07	TXT File	1 KB
meta_linear_cna.txt	25/03/2022 19:07	TXT File	1 KB
meta_methylation_hm450.txt	25/03/2022 19:07	TXT File	1 KB
meta_mrna_seq_v2_rsem.txt	25/03/2022 19:07	TXT File	1 KB
meta_mrna_seq_v2_rsem_zscores_ref_all_samples.txt	25/03/2022 19:07	TXT File	1 KB
meta_mrna_seq_v2_rsem_zscores_ref_diploid_samples.txt	25/03/2022 19:07	TXT File	1 KB
meta_mutations.txt	25/03/2022 19:07	TXT File	1 KB
meta_rppa.txt	25/03/2022 19:07	TXT File	1 KB
meta_rppa_zscores.txt	25/03/2022 19:07	TXT File	1 KB
meta_study.txt	25/03/2022 19:07	TXT File	1 KB

For each assay, 1 data file and 1 metadata/information file

Understanding the data

→  data_clinical_patient.txt	25/03/2022 19:15	TXT File	355 KB
 data_clinical_sample.txt	25/03/2022 19:15	TXT File	103 KB

TSV – feature x patient ID (many missing values, cancer-specific features)

Patient information

Sex, height, weight, race, ethnicity, diagnosis age, survival status


Occupation history, smoking status, family history

Tumour information

Stage, grade, disease codes, metastasis status

Tumour-specific categories (e.g. for bladder, rate of prostate cancer)

Understanding the data

 data_cna.txt

25/03/2022 19:15




TXT File

22,499 KB

tumour x gene using GISTIC scale (TSV)

- 2 homozygous “deep” deletion
- 1 shallow deletion (anything that isn’t total loss)
- 0 diploid
- 1 gain (“a few” extra copies)
- 2 amplification (often in focal sets)


Understanding the data

→	 data_mrna_seq_v2_rsem.txt	25/03/2022 19:15	TXT File	69,336 KB
	 data_mrna_seq_v2_rsem_zscores_ref_all_samples.txt	25/03/2022 19:15	TXT File	60,450 KB
	 data_mrna_seq_v2_rsem_zscores_ref_diploid_samples.txt	25/03/2022 19:15	TXT File	59,941 KB

tumour x gene, normalised gene expression data (TSV)

- Normalised counts
- Can feed into differential expression pipelines (DESeq2 *etc*), if careful!
- Good for comparisons of one gene across samples
- Harder to compare expression between genes of same sample

Understanding the data

 data_mutations.txt

25/03/2022 19:15

TXT File

264,752 KB ←

TSV – list of all mutations, sorted by tumour ID

- Includes synonymous mutations as well as non-synonymous
- Data structure is rubbish, requires lots of parsing to find hotspots *etc.*

Working with the data



Existing UG training and extensive core bioinformatic support



[Specific cBioPortal REST API for programmatic access](#)



Python support available too – pandas package is versatile



Doable...! But. Memory intensive, and watch delimiters when importing.

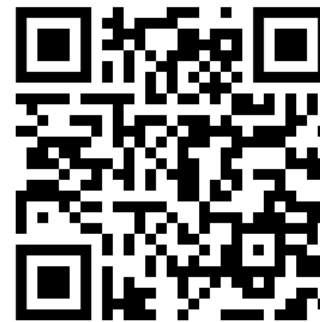
cBioPortal API demonstration using RStudio

Introduction to cBioPortal

Course material



Google Form for questions



Working with the data... final thoughts

The data is not always complete

- Inconsistent column usage between datasets
- Watch 'whitespace' vs 'tab space' vs comma delimiters

Biological vs Statistical significance

Limited by previous bioinformatic analysis pipelines, genome version *etc.*

- More advanced questions can go back to the raw data



Course complete!