AACR American Association for Cancer Research®
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10
#AACR24
AACR.ORG/AACR24

A clinico-genomic data processing pipeline using the {genieBPC} R package

Samantha Brown
Katherine S. Panageas
Memorial Sloan Kettering Cancer Center, New York, NY

Disclosure Information

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

Samantha Brown

I have the following relevant financial relationships to disclose:

- Received salary support from AACR Project GENIE Biopharma Collaborative (BPC).
- Received support from NCI Cancer Center Support Grant P30 CA008748.

Katherine S. Panageas

I have the following relevant financial relationships to disclose:

- Received salary support from AACR Project GENIE Biopharma Collaborative (BPC).
- Received support from NCI Cancer Center Support Grant P30 CA008748.

Agenda

Projects GENIE & GENIE BPC

Clinico-Genomic Data Processing Pipeline

Case study

Data processing with {genieBPC}

Conclusion

American Association for Cancer Research Project GENIE

AACR American Association for Cancer Research
ANNUAL MEETING 2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

- AACR **Project GENIE** (Genomics Evidence Neoplasia Information Exchange) is a publicly accessible international cancer registry of genomic data assembled through data sharing agreements between 19 of the leading cancer centers in the world
  - GENIE includes genomic data from targeted sequencing panels and limited clinical data (age, sex, date of diagnosis, cancer type and date of death)
  - Genomic data for 195,000 samples is currently available

AACR American Association for Cancer Research
FINDING CURES TOGETHER

PROJECT GENIE
Genomics Evidence Neoplasia Information Exchange

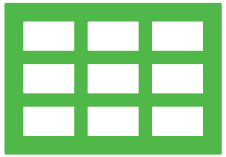
https://www.aacr.org/professionals/research/aacr-project-genie/

AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. 2017 Aug;7(8):818-831. doi: 10.1158/2159-8290.CD-17-0151. Epub 2017 Jun 1. PMID: 28572459; PMCID: PMC5611790.

Projects GENIE & GENIE BPC

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

The goal of **Project GENIE Biopharma Collaborative (BPC)** is to augment the existing registry genomic data from AACR Project GENIE with enhanced clinical (phenomic) data to support clinical-genomics analyses.



Phenomic data are curated using the PRISSMM curation model to capture detailed information on cancer diagnosis, drug regimens, disease status from radiology reports, pathology reports and medical oncologist assessments, structured in several datasets with over 700 feature variables.

Analyses using linked clinico-genomic databases – including GENIE BPC – will help to drive advancements in precision oncology in identifying the genomic alterations and drug therapies that optimize clinical outcomes.

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

PRISSMM™:
A Taxonomy for Defining Cancer Outcomes

Pathologic evidence of locoregional or distant evidence of tumor

Radiographic evidence of locoregional recurrent or persistent tumor

Imaging evidence of distant/disseminated tumor beyond the primary site

Symptoms of tumor on physical exam or symptoms that can be attributed to tumor

Signs of cancer on physical exam or symptoms that can be attributed to tumor

Tumor **M**arker evidence of persistent or recurrent tumor

Oncology **M**edical Provider assessment

Each curation effort may focus on some or all of the PRISSMM™ components

Note: neither RECIST data nor patient reported outcomes data are captured via PRISSMM.

GENIE Biopharma Collaborative

AACR American Association for Cancer Research
ANNUAL MEETING 2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

- Data includes patients with ≥1 high-throughput sequencing profile
- Four participating institutions for Phase I: Currently Memorial Sloan Kettering, Dana Farber, Vanderbilt and University Health Network

Cancer Cohort	N	Status
Non-small cell lung cancer	1832	Publicly available
Colorectal cancer	1479	Publicly available
Breast cancer*	1130	Data currently available to consortium members
Pancreas cancer*	1109	Data currently available to consortium members
Prostate cancer*	1116	Data currently available to consortium members
Bladder cancer*	716	Data currently available to consortium members

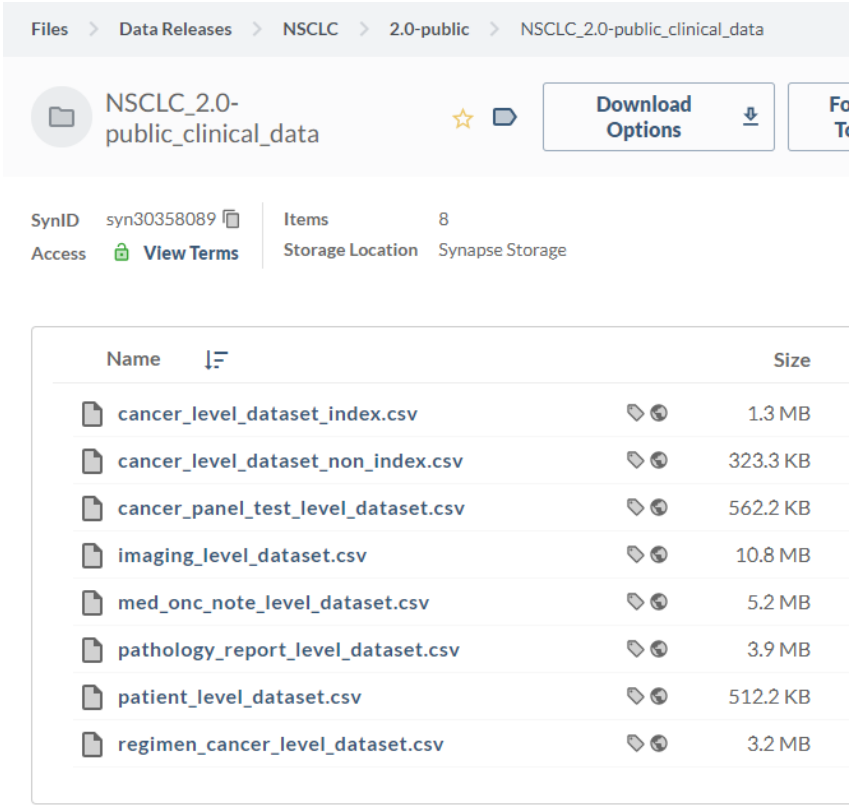
Cancer Cohort	N	Status
Non-small cell lung cancer, additional cases	1717	Undergoing quality assurance processes
Colorectal cancer, additional cases	1481	Undergoing quality assurance processes
Renal cell carcinoma	1302	Beginning curation
Ovarian cancer	1294	Testing data dictionary
Esophagogastric cancer	1297	Planned 2025 data release
Melanoma	1294	Planned 2025 data release

*Data to be publicly released in 2024

GENIE BPC Data

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

- Data are publicly released by cancer cohort
  - In phase I: non-small cell lung (NSCLC), colorectal (CRC), breast, pancreas, prostate, bladder
- New versions of data are released periodically to include additional patients and variables and to incorporate data corrections
- *.csv* and *.txt* data files are available for download from Sage Bionetworks' Synapse data sharing platform
  - For each data release, an Analytic Data Guide that defines each variable is available and should be referenced
- Downloading each file individually poses challenges for efficient and reproducible workflows



{genieBPC} R Package

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

The {genieBPC} package is a pipeline to programmatically access the data corresponding to each release from Synapse to support reproducibility, and to create datasets linking clinical and genomic data for analysis.

Created and developed by

Samantha Brown

Michael Curry

Hannah Fuchs

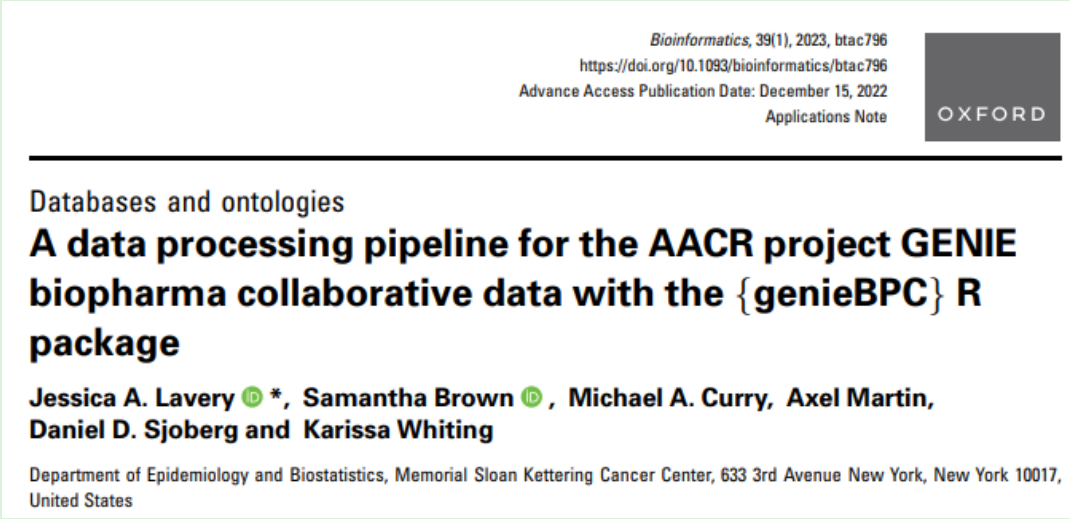
Jessica Lavery

Axel Martin

Dan Sjoberg

Karissa Whiting

{genieBPC} Publication



Bioinformatics, 39(1), 2023, btac796
https://doi.org/10.1093/bioinformatics/btac796
Advance Access Publication Date: December 15, 2022
Applications Note

OXFORD

Databases and ontologies

A data processing pipeline for the AACR project GENIE biopharma collaborative data with the {genieBPC} R package

Jessica A. Lavery *, Samantha Brown, Michael A. Curry, Axel Martin, Daniel D. Sjoberg and Karissa Whiting

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 633 3rd Avenue New York, New York 10017, United States

Jessica A Lavery, Samantha Brown, Michael A Curry, Axel Martin, Daniel D Sjoberg, Karissa Whiting, A data processing pipeline for the AACR project GENIE biopharma collaborative data with the {genieBPC} R package, *Bioinformatics*, Volume 39, Issue 1, January 2023, btac796, https://doi.org/10.1093/bioinformatics/btac796

Register for a Synapse Account

Instructions:

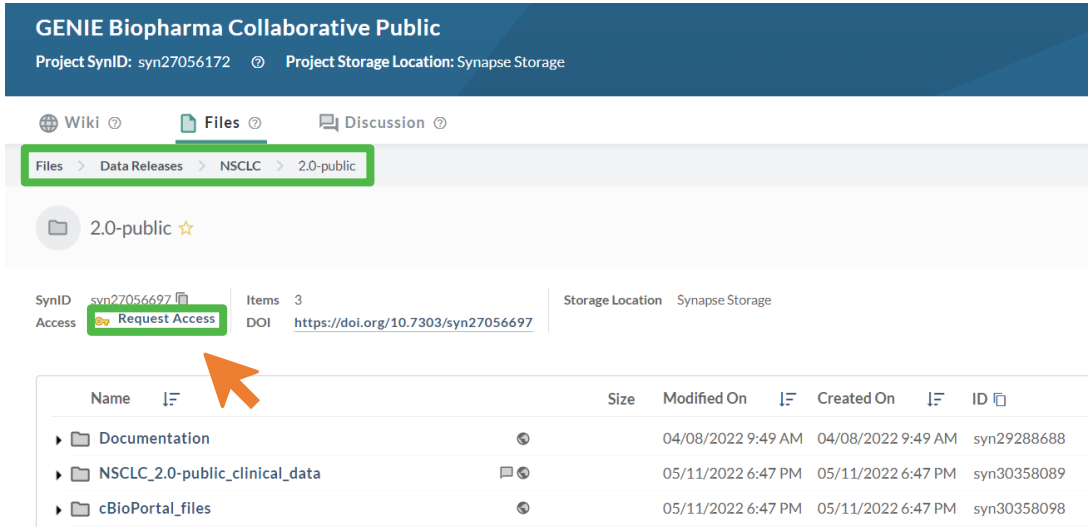
1. Register for a 'Synapse' account. Be sure to create a username and password. **Do NOT connect via your Google account.**

   https://www.synapse.org/#

2. Accept the **Synapse account terms of use.**
3. Navigate to GENIE Biopharma Collaborative Public page

   https://www.synapse.org/#!Synapse:syn27056172/wiki/616601

4. In the Files folder, navigate to Data Releases -> NSCLC -> 2.0-public
5. Select *Request Access*, review the **terms of data use** and click *Accept*



Installation Instructions

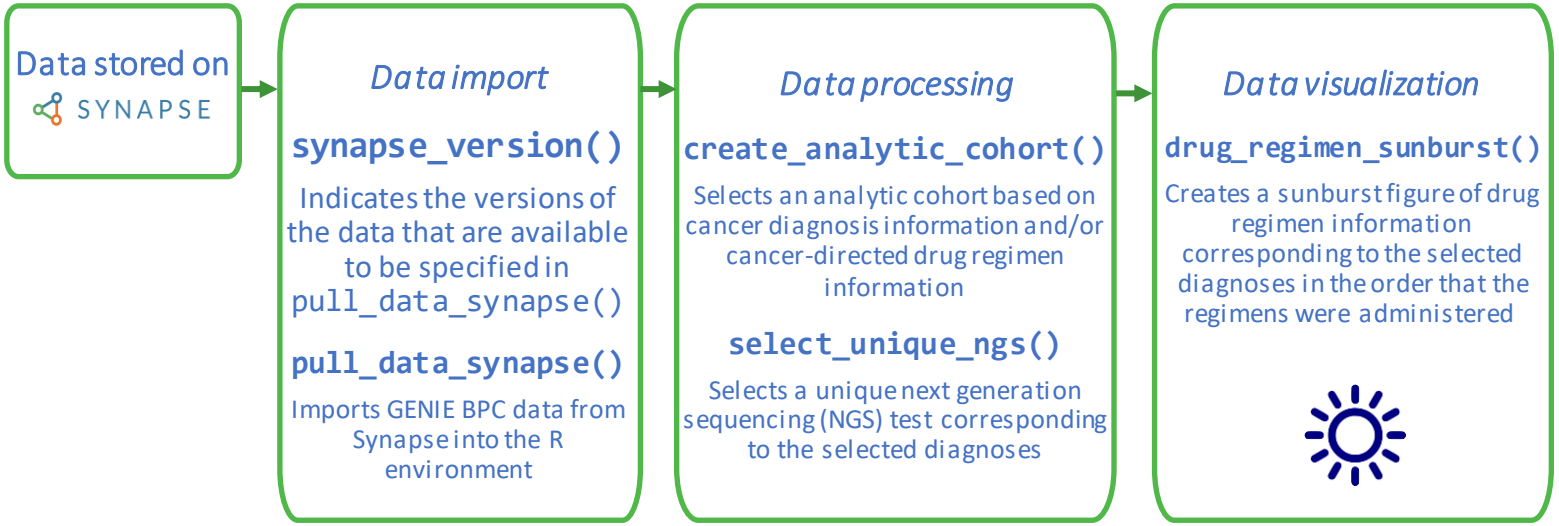
Installing {genieBPC}:

```
install.packages("genieBPC")
```

- These instructions are also included in the Demo.R script on our GitHub repository: https://github.com/GENIE-BPC/intro_to_genieBPC
- Further R package details are available on the {genieBPC} GitHub repo & website
- {genieBPC} requires R version >=3.6

Clinico-Genomic Data Processing Pipeline

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24



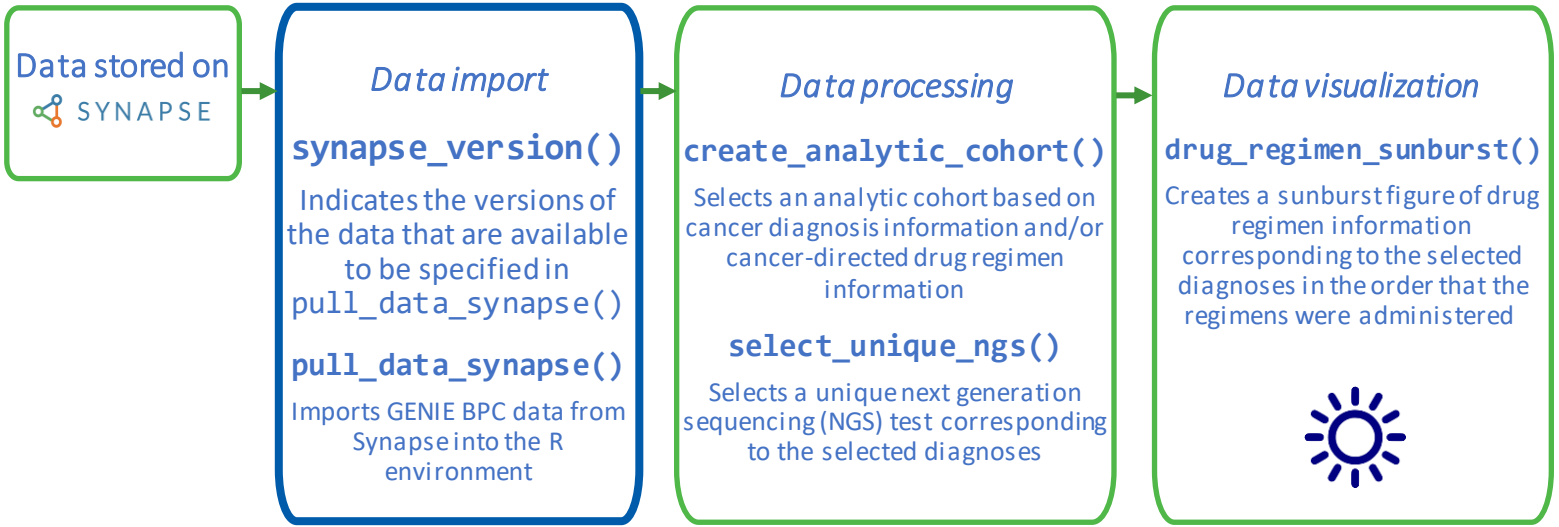
Case Study

Create a cohort of patients who were diagnosed with Stage IV adenocarcinoma non-small cell lung cancer (NSCLC) and received Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab as their first cancer-directed drug regimen after diagnosis.

Follow along using the Demo.R script on our GitHub repository: https://github.com/GENIE-BPC/intro_to_genieBPC

Clinico-Genomic Data Processing Pipeline

AACR American Association for Cancer Research
ANNUAL MEETING 2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24



Set Synapse Credentials

To pull data from Synapse, users must create a Synapse account and store their Synapse credentials in the R environment. The **`set_synapse_credentials()`** function will store credentials during each R session:

```
set_synapse_credentials(username = 'your_username',
                        password = 'your_password')
```

Set Synapse Credentials

To pull data from Synapse, users must create a Synapse account and store their Synapse credentials in the R environment. The **`set_synapse_credentials()`** function will store credentials during each R session:

```
set_synapse_credentials(username = 'your_username',
                        password = 'your_password')
```

Future enhancement

Additional functionality will be released soon to allow users to pass their Synapse Personal Access Token (PAT) through **`set_synapse_credentials()`**:

```
set_synapse_credentials(pat = 'your_pat')
```

`synapse_version()`

- Helper function that returns a table of GENIE BPC data releases that are currently available
- `synapse_version()` input parameter: most_recent = TRUE/FALSE
  - Calling `genieBPC::synapse_version(most_recent = TRUE)` will return a table with each cancer cohort and its latest data release version
  - Calling `genieBPC::synapse_version(most_recent = FALSE)` will return a table with all cancer cohorts and data releases available

`synapse_version()`

- Helper function that returns a table of GENIE BPC data releases that are currently available
- `synapse_version()` input parameter: most_recent = TRUE/FALSE
  - Calling `genieBPC::synapse_version(most_recent = TRUE)` will return a table with each cancer cohort and its latest data release version
  - Calling `genieBPC::synapse_version(most_recent = FALSE)` will return a table with all cancer cohorts and data releases available

`synapse_version(most_recent = TRUE)`

cohort	version	release_date	all_versions
BLADDER	v1.2-consortium	November 2023	Most Recent Versions
BrCa	v1.2-consortium	October 2022	Most Recent Versions
CRC	v1.3-consortium	February 2024	Most Recent Versions
CRC	**v2.0-public**	**October 2022**	**Most Recent Versions**
NSCLC	v2.2-consortium	February 2024	Most Recent Versions
NSCLC	**v2.0-public**	**May 2022**	**Most Recent Versions**
PANC	v1.2-consortium	January 2023	Most Recent Versions
Prostate	v1.2-consortium	January 2023	Most Recent Versions

pull_data_synapse()

- Pull GENIE BPC clinical and genomic data directly from Synapse into R
- Can specify cancer type (`cohort`) and version of data (`version`)
  - Version of the data is updated periodically on Synapse with re-releases (new variables available, additional QA, etc.)
- Returns a nested list of data frames for each cancer site for the accompanying version

Argument	Description	Acceptable Values
`cohort`	• GENIE BPC Project cancer • Currently, NSCLC and CRC are the only two publicly available datasets	• NSCLC • CRC • BrCa • PANC • Prostate • BLADDER
`version`	Version of the data (e.g v1.1-consortium, v2.0-public)	• Values can be found in `synapse_version()`

Demo: Run `pull_data_synapse()` for case study

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
library(genieBPC)

set_synapse_credentials()

nsclc_synapse_data <-
    pull_data_synapse(
      cohort = "NSCLC",
      version = "v2.0-public")
```

Demo: Run `pull_data_synapse()` for case study

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
library(genieBPC)

set_synapse_credentials()

nsclc_synapse_data <-
    pull_data_synapse(
      cohort = "NSCLC",
      version = "v2.0-public")
```

- pt_char
- ca_dx_index
- ca_dx_non_index
- ca_drugs
- prissmm_pathology
- prissmm_imaging
- prissmm_md
- cpt
- mutations_extended
- cna
- fusions

Clinico-Genomic Data Processing Pipeline

AACR American Association for Cancer Research
ANNUAL MEETING 2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

Data stored on SYNAPSE

Data import

`synapse_version()`

Indicates the versions of the data that are available to be specified in `pull_data_synapse()`

`pull_data_synapse()`

Imports GENIE BPC data from Synapse into the R environment

Data processing

`create_analytic_cohort()`

Selects an analytic cohort based on cancer diagnosis information and/or cancer-directed drug regimen information

`select_unique_ngs()`

Selects a unique next generation sequencing (NGS) test corresponding to the selected diagnoses

Data visualization

`drug_regimen_sunburst()`

Creates a sunburst figure of drug regimen information corresponding to the selected diagnoses in the order that the regimens were administered

create_analytic_cohort()

Create a cohort from the GENIE BPC data

Cancer diagnosis information such as cancer cohort, treating institution, histology, and stage at diagnosis

Cancer-directed regimen information including regimen name and regimen order.

This function returns all clinical and genomic data for the selected patients

GENIE BPC Clinical Datasets

Patient characteristics
1 row /patient

Cancer diagnosis
1 row /cancer diagnosis

Cancer-directed drugs
1 row /drug regimen/associated cancer dx

PRISSMM Imaging
1 row /imaging report

PRISSMM Pathology
1 row /pathology report

PRISSMM Medical Oncologist Assessments
1 row /med onc assessment

PRISSMM Tumor Marker Assessments
1 row /tumor marker result

Cancer Panel Test
1 row /CPT report/associated cancer dx

create_analytic_cohort()

Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	• Name of object in global environment that was returned from pull_data_synapse()

create_analytic_cohort()

Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	• Name of object in global environment that was returned from pull_data_synapse()
index_ca_seq	Index cancer sequence. Default is 1, indicating the patient's first index cancer. This refers to the cancer with associated genomic sequencing.	• Numeric (1+)

create_analytic_cohort()

Argument	Description	Acceptable Values
`data_synapse`	List returned from `pull_data_synapse()`	• Name of object in global environment that was returned from `pull_data_synapse()`
`index_ca_seq`	Index cancer sequence. Default is 1, indicating the patient's first index cancer. This refers to the cancer with associated genomic sequencing.	• Numeric (1+)
`institution`	GENIE BPC participating institution. Default selection is all institutions. *Note that not all institutions curated data for all cancer sites.*	• DFCI • MSK • UHN • VICC

create_analytic_cohort()

Argument	Description	Acceptable Values
`stage_dx`	Stage at diagnosis. Default selection is all stages.	• Stage I • Stage II • Stage III • Stage I-III NOS • Stage IV
`histology`	Cancer histology. Default selection is all histologies. For all cancer cohorts except for BrCa (breast cancer), this parameter corresponds to the variable 'ca_hist_adeno_squamous'. For BrCa, this parameter corresponds to the variable 'ca_hist_brca'	All cancer types except breast: • Adenocarcinoma • Squamous cell • Sarcoma • Small cell carcinoma • Other histologies/mixed tumor Breast cancer: • Invasive lobular carcinoma • Invasive ductal carcinoma • Other histology

create_analytic_cohort()

Argument	Description	Acceptable Values
`regimen_drugs`	Vector with names of drugs in cancer-directed regimen, separated by a comma. For example, to specify a regimen consisting of Carboplatin and Pemetrexed Disodium, specify `regimen_drugs` = "Carboplatin, Pemetrexed Disodium".	Acceptable values are found in the `drug_names_by_cohort` dataset provided with this package.
`regimen_type`	Indicates whether the regimen(s) specified in `regimen_drugs` indicates the exact regimen to return, or if regimens containing the drugs listed in `regimen_drugs` should be returned.	• Exact • Containing

Example: `regimen_drugs` and `regimen_type`

`regimen_drugs`	`regimen_type`	Example regimens returned
Carboplatin	Exact	• Carboplatin
Carboplatin	Containing	• Carboplatin • Carboplatin, Cisplatin • Carboplatin, Paclitaxel • Carboplatin, Pemetrexed Disodium • etc.

create_analytic_cohort()

Argument	Description	Acceptable Values
`regimen_order`	Order of cancer-directed regimen. If multiple drugs are specified, `regimen_order` indicates the regimen order for all drugs; different values of `regimen_order` cannot be specified for different drug regimens.	• Numeric (1+)
`regimen_order_type`	Specifies whether the 'regimen_order' parameter refers to the order of receipt of the drug regimen within the cancer diagnosis (across all other drug regimens; "within cancer") or the order of receipt of the drug regimen within the times that that drug regimen was administered (“within regimen”)	• Within cancer • Within regimen

Example: `regimen_order` and `regimen_order_type`

regimen_drugs	regimen_order	regimen_order_type	Specified output
Carboplatin, Pemetrexed Disodium	1	Within cancer	The first time that Carboplatin + Pemetrexed was received, among all drug regimens associated with that cancer diagnosis
Carboplatin, Pemetrexed Disodium	1	Within regimen	The first instance that Carboplatin + Pemetrexed was received, out of all times that the patient received Carboplatin + Pemetrexed

create_analytic_cohort()

Argument	Description	Acceptable Values
`return_summary`	Specifies whether summary tables are returned using {gtsummary}. Default is FALSE.	• TRUE • FALSE

Demo: `create_analytic_cohort()` for case study using NSCLC 2.0-public data

American Association for Cancer Research®
FINDING CURES TOGETHER®

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

Demo: `create_analytic_cohort()` for case study using NSCLC 2.0-public data

American Association for Cancer Research
FINDING CURES TOGETHER

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
nsclc_cohort <- create_analytic_cohort(
  data_synapse = nsclc_synapse_data$NSCLC_v2.0,
  stage_dx = c("Stage IV"),
  histology = "Adenocarcinoma",
  regimen_drugs = c("Carboplatin, Pemetrexed Disodium",
                    "Cisplatin, Pemetrexed Disodium",
                    "Bevacizumab, Carboplatin, Pemetrexed Disodium",
                    "Bevacizumab, Cisplatin, Pemetrexed Disodium"),
  regimen_type = "Exact",
  regimen_order = 1,
  regimen_order_type = "within cancer",
  return_summary = TRUE
)
```

Demo: `create_analytic_cohort()` for case study using NSCLC 2.0-public data

American Association for Cancer Research
FINDING CURES TOGETHER

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

Calling `nsclc_cohort` returns a list of datasets:

- cohort_pt_char
- cohort_ca_dx_index
- cohort_ca_dx_non_index
- cohort_ca_drugs
- cohort_prissmm_pathology
- cohort_prissmm_imaging
- cohort_prissmm_md
- cohort_cpt
- cohort_mutations_extended
- cohort_cna
- cohort_fusions

Additionally, the list contains summary table objects when `return_summary = TRUE`:

- tbl_overall_summary
- tbl_cohort
- tbl_drugs
- tbl_ngs

```
nsclc_cohort$
tbl_overall_summary
```

Overall Summary

Characteristic	N = 241 patients[1]
Number of diagnoses per patient in cohort_ca_dx data frame	
1	241 (100%)
Number of regimens per patient in cohort_ca_drugs data frame	
1	241 (100%)
Number of CPTs per patient in cohort_ngs data frame	
1	222 (92%)
2	18 (7.5%)
4	1 (0.4%)

[1] n (%)

```
nsclc_cohort$
tbl_cohort
```

Cohort Summary

Characteristic	N = 241 Diagnoses[1]
Cohort (cohort)	
NSCLC	241 (100%)
Institution (institution)	
DFCI	92 (38%)
MSK	118 (49%)
VICC	31 (13%)
Stage at diagnosis (stage_dx)	
Stage IV	241 (100%)
Histology (ca_hist_adeno_squamous)	
Adenocarcinoma	241 (100%)

[1] n (%)

```
nsclc_cohort$
tbl_drugs
```

Cancer-Directed Drugs Summary

Characteristic	N = 241 Regimens[1]
Cohort (cohort)	
NSCLC	241 (100%)
Institution (institution)	
DFCI	92 (38%)
MSK	118 (49%)
VICC	31 (13%)
Drugs in regimen (regimen_drugs)	
Bevacizumab, Carboplatin, Pemetrexed Disodium	52 (22%)
Bevacizumab, Cisplatin, Pemetrexed Disodium	27 (11%)
Carboplatin, Pemetrexed Disodium	124 (51%)
Cisplatin, Pemetrexed Disodium	38 (16%)

[1] n (%)

```
nsclc_cohort$
tbl_ngs
```

NGS Summary

Characteristic	N = 262 Cancer Panel Tests[1]
Cohort (cohort)	
NSCLC	262 (100%)
Institution (institution)	
DFCI	99 (38%)
MSK	126 (48%)
VICC	37 (14%)
OncoTree code (cpt_oncotree_code)	
LCLC	1 (0.4%)
LUAD	253 (97%)
LUAS	1 (0.4%)
LUSC	1 (0.4%)
NSCLC	4 (1.5%)
NSCLCPD	2 (0.8%)
Sequence assay ID (cpt_seq_assay_id)	
DFCI-ONCOPANEL-1	1 (0.4%)
DFCI-ONCOPANEL-2	57 (22%)
DFCI-ONCOPANEL-3	41 (16%)
MSK-IMPACT341	3 (1.1%)
MSK-IMPACT410	61 (23%)
MSK-IMPACT468	62 (24%)
VICC-01-SOLIDTUMOR	26 (9.9%)
VICC-01-T5A	1 (0.4%)
VICC-01-T7	10 (3.8%)

[1] n (%)

`select_unique_ngs()`

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

Selecting one genomic sample per patient:

While patients can have many NGS reports, we often need to select a single sample per patient for analyses.

The **`select_unique_ngs()`** function selects one sample per patient.

This function prioritizes characteristics of interest (e.g., sample type).

Note: if a patient only has one report, it will be returned regardless of criteria.

After running `select_unique_ngs()`, the user will be ready to process the genomic data. The `{gnomeR}` R package contains many tools to facilitate annotation and analysis of complex genomic data.

See https://mskcc-epi-bio.github.io/gnomeR/ for more details.

select_unique_ngs()

Argument	Description	Acceptable Values
`data_cohort`	Output object of the `create_analytic_cohort()` function	• Name of NGS object in global environment that was returned from `create_analytic_cohort()`
`oncotree_code`	Character vector specifying which sample OncoTree codes to prioritize.	• See 'cpt_oncotree_code' column of data_cohort.
`sample_type`	Character specifying which type of genomic sample to prioritize. Options are 'Primary', 'Local', and 'Metastasis'. Default is to not select a NGS sample based on the sample type.	• Primary • Local • Metastasis
`min_max_time`	Character specifying if the first or last genomic sample recorded should be kept.	• min (refers to earliest sample) • max (refers to latest sample)

Demo: `select_unique_ngs()` for case study using NSCLC 2.0-public data

American Association for Cancer Research
FINDING CURES TOGETHER

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
nrow(nsclc_cohort$cohort_ngs)
```

[1] 262

```
nsclc_samp <- select_unique_ngs(
   data_cohort = nsclc_cohort$cohort_ngs,
   oncotree_code = "LUAD",
   sample_type = "Metastasis",
   min_max_time = "max")

nrow(nsclc_samp)
```

[1] 241

{gnomeR}

The {gnomeR} package provides a consistent framework for genetic data wrangling, processing, visualization, and analysis.

Wrangling

- Addresses issues faced when processing multi-institutional genomic data, for example:
  - Accounting for various gene panels
  - Inconsistent data formats and gene standards

Processing

- `create_gene_binary()` function processes mutation, fusions, and CNA data into analytic format
- `summarize_by_gene()` function allows users to analyze on the gene level instead of the alteration level

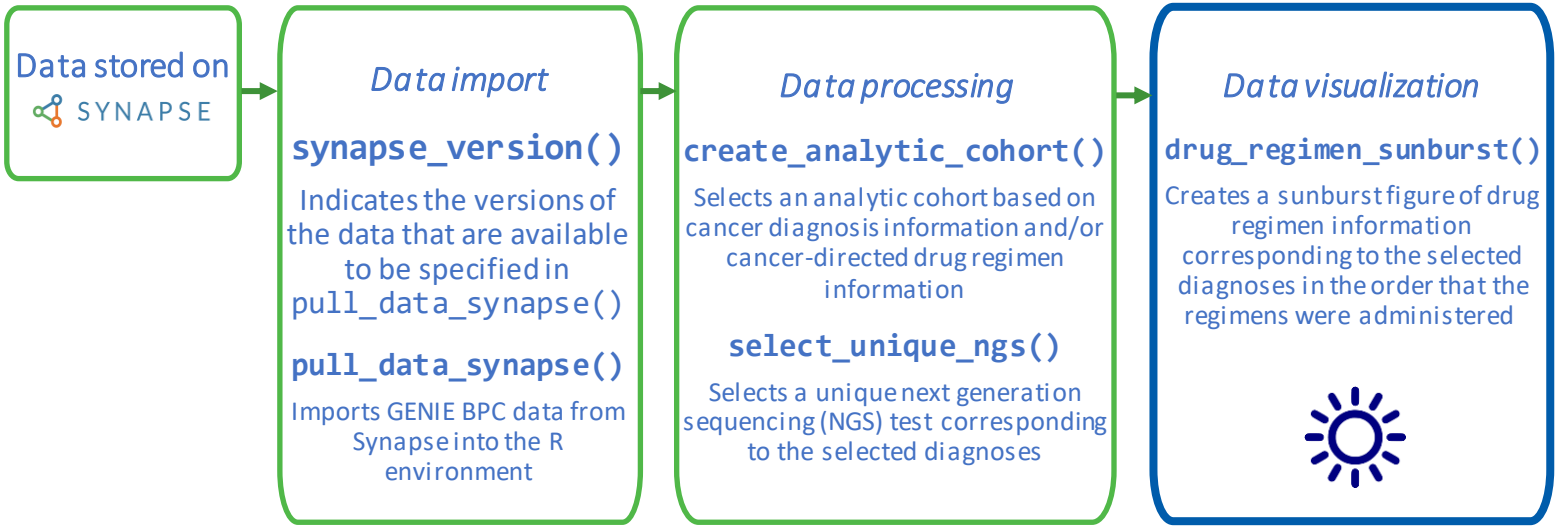
Visualization

- `ggcomut()` function creates a co-mutation heatmap of most frequently altered genes
- `ggtopgenes()` function creates a barchart of most frequently altered genes

See https://mskcc-epi-bio.github.io/gnomeR/ for more details.

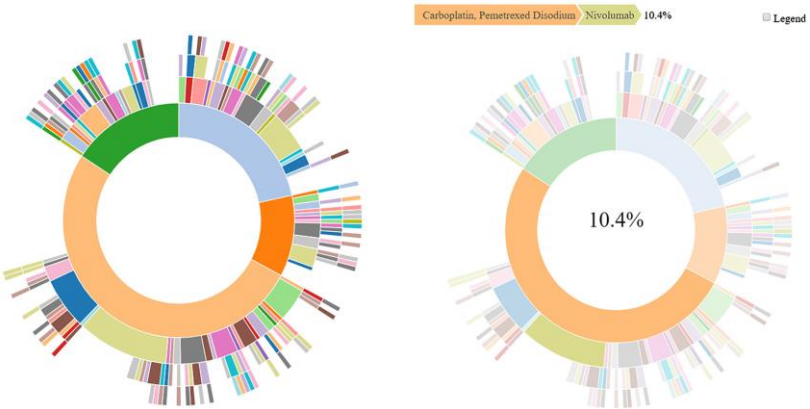
Clinico-Genomic Data Processing Pipeline

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24



drug_regimen_sunburst()

- Visualize the complete treatment course for selected cancer diagnoses
- Each ring corresponds to a regimen (i.e., innermost ring is first regimen, second innermost ring is second regimen, etc.)
- Interactive figure: Can hover to see regimen names and percent of patients receiving that regimen



drug_regimen_sunburst()

Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	• Name of object in global environment that was returned from pull_data_synapse()
data_cohort	The list returned from the create_analytic_cohort() function call	• Name of object in global environment that was returned from create_analytic_cohort()
max_n_regimens	The maximum number of regimens displayed in the sunburst plot	• Integer >0

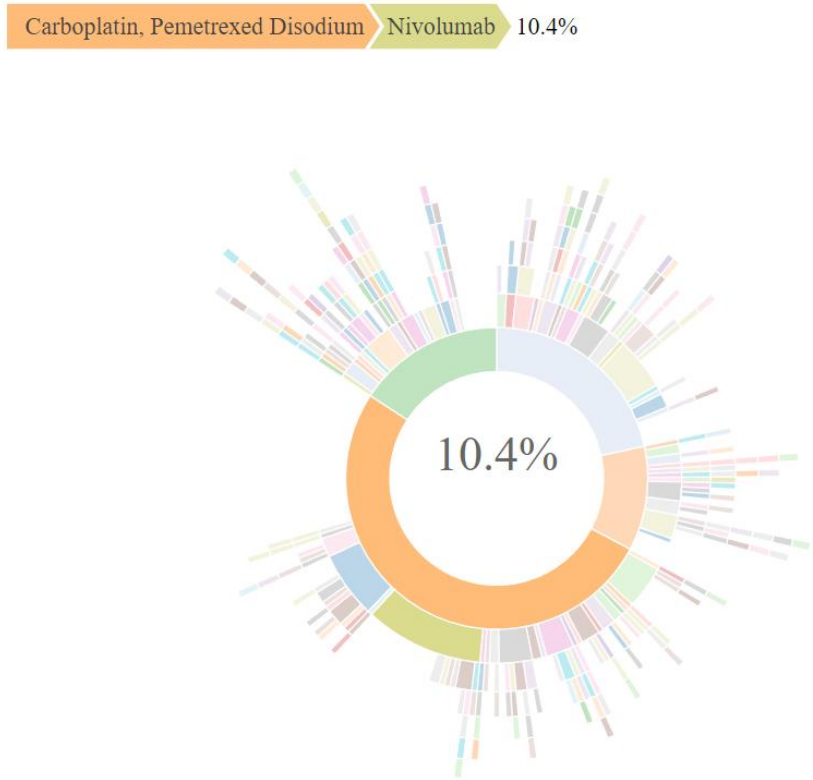
Demo: `drug_regimen_sunburst()` for case study using NSCLC 2.0-public data

American Association for Cancer Research®
FINDING CURES TOGETHER®

Case Study: Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
nsclc_sunburst <- drug_regimen_sunburst(
          data_synapse = nsclc_synapse_data$NSCLC_v2.0,
          data_cohort = nsclc_cohort)
```

```
nsclc_sunburst$
sunburst_plot
```



Future {genieBPC} R Package Enhancements

- Selection of multiple cohorts simultaneously (a single call to `create_analytic_cohort()` instead of multiple calls in order to pull patients across cancer types based on similar criteria)
- Cohort selection based on sites of metastatic disease
- Access to Synapse via Personal Access Token (PAT), in addition to username and password

Suggestions? File an issue on GitHub
https://github.com/GENIE-BPC/genieBPC/issues

The Future of Project GENIE BPC

AACR American Association for Cancer Research
ANNUAL MEETING
2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

- Currently onboarding additional participating institutions
- Curation of additional cancer sites and additional cases for existing cancer sites
  - Additional NSCLC, CRC cases
  - Renal cell carcinoma
  - Ovarian cancer
  - Esophagogastric cancer
  - Melanoma

Conclusion

The **{genieBPC}** R package offers a reproducible pipeline to create cohorts for clinico-genomic analyses by streamlining data access and clinical data processing from multiple clinical data files of varying structure to create analytic cohorts.

Thank you!

Thank you to Jessica Lavery, Hannah Fuchs, and Karissa Whiting for contributions to slides, and to contributing {genieBPC} authors: Michael Curry, Hannah Fuchs, Axel Martin, Dan Sjoberg, Karissa Whiting

Project GENIE BPC Acknowledgements

AACR American Association for Cancer Research
ANNUAL MEETING 2024 • SAN DIEGO
APRIL 5-10 • AACR.ORG/AACR24 • #AACR24

AACR GENIE Coordinating Center
Shawn Sweeney
Alyssa Acebedo
Mike Fiandalo

Dana Farber Cancer Institute
Ken Kehl
Asha Postle
Bill Hahn
John Orechia

Vanderbilt
Rhonda Potter
Christine Micheel
Sandip Chaugai
Ben Ho Park
Sanjay Mishra
Lucy Wang

University Health Network
Celeste Yu
Philippe Bedard

Memorial Sloan Kettering
Charles Sawyers
Gregory Riely
Deb Schrag
Julia Rudolph
Chelsea Nichols
Shirin Pillai
John Phillip
Marufur Bhuiya
Stu Gardos
Cynthia Chu
Rona Yaeger
Pedram Razavi
Anna Varghese
Wassim Abida
David Jones
Ronglai Shen
Yuan Chen
Karissa Whiting

Sage Bionetworks
Xindi Guo
Chelsea Nayan
Thomas Yu

VASTA Global
Melanie Bernstein

cBioPortal
Niki Schultz
Ritika Kundra
Brooke Mastrogiacomo
Ino de Bruijn

Statistical Coordinating Center
Kathy Panageas
Jessica Lavery
Samantha Brown
Hannah Fuchs