Analyzing Clinical and Genomic Oncological Data with {genieBPC} and {gnomeR}

GENIE BPC Winter Symposium January 16, 2024

Sammi Brown

Hannah Fuchs

Jessica Lavery

Karissa Whiting





Agenda



Clinico-Genomic Data Processing Pipeline



Case study



Clinical data processing with {genieBPC}



Genomic data processing with {gnomeR}



Conclusion

{genieBPC} & {gnomeR} R Packages



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



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



Created and developed by

Arshi Arora

Michael Curry

Hannah Fuchs

Axel Martin

Karissa Whiting

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_and_gnomeR
- Further R package details are available on the {genieBPC} <u>GitHub repo</u> & <u>website</u>
- {genieBPC} requires R version >= 3.6

Clinico-Genomic Data Processing Pipeline

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



Genomic Processing

{gnomeR}

create_gene_binary()

Processes data on mutation, CNA and fusions into analytic format

tbl_genomic()

Summarizes gene alterations across clinical variables of interest

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 and gnomeR

Clinico-Genomic Data Processing Pipeline

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



Genomic Processing

{gnomeR}

create_gene_binary()

Processes data on mutation, CNA and fusions into analytic format

tbl_genomic()

Summarizes gene alterations across clinical variables of interest

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

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:

Coming soon

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() has one input: 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() has one input: 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	v1.2-consortium	August 2021	Most Recent Versions
CRC	v2.0-public	October 2022	Most Recent Versions
NSCLC	v2.1-consortium	August 2021	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()





library(genieBPC)





library(genieBPC)
set_synapse_credentials()





```
library(genieBPC)
set_synapse_credentials()
nsclc_synapse_data <- pull_data_synapse(cohort = "NSCLC", version = "v2.0-public")</pre>
```





```
library(genieBPC)
set_synapse_credentials()
nsclc_synapse_data <- pull_data_synapse(cohort = "NSCLC", version = "v2.0-public")
Calling nsclc_synapse_data$NSCLC_v2.0 returns a list of datasets in nsclc_synapse_data:</pre>
```

- 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

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



Genomic Processing

{gnomeR}

create_gene_binary()

Processes data on mutation, CNA and fusions into analytic format

tbl_genomic()

Summarizes gene alterations across clinical variables of interest



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

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()

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 is the cancer that met the eligibility criteria for the project and was selected at random for PRISSMM phenomic data curation.	Numeric (1+)

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 is the cancer that met the eligibility criteria for the project and was selected at random for PRISSMM phenomic data curation.	• Numeric (1+)
institution	GENIE BPC participating institution. Default selection is all institutions. Note that not all institutions curated data for all cancer sites.	DFCIMSKUHNVICC

Argument	Description	Acceptable Values
stage_dx	Stage at diagnosis. Default selection is all stages.	 Stage I Stage III Stage I-III NOS Stage IV

Argument	Description	Acceptable Values
stage_dx	Stage at diagnosis. Default selection is all stages.	 Stage I 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 carcinomaOther histology

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.	ExactContaining

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.

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 (e.g. the first time carboplatin pemetrexed was received, out of all times that the patient received carboplatin pemetrexed; "within regimen").	Within cancerWithin regimen

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





nsclc_cohort <- create_analytic_cohort(</pre>





```
nsclc_cohort <- create_analytic_cohort(
  data_synapse = nsclc_synapse_data$NSCLC_v2.0,</pre>
```





```
nsclc_cohort <- create_analytic_cohort(
  data_synapse = nsclc_synapse_data$NSCLC_v2.0,
  stage_dx = c("Stage IV"),</pre>
```





```
nsclc_cohort <- create_analytic_cohort(
  data_synapse = nsclc_synapse_data$NSCLC_v2.0,
  stage_dx = c("Stage IV"),
  histology = "Adenocarcinoma",</pre>
```





```
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",</pre>
```





















```
nsclc cohort <- create analytic cohort(</pre>
  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,
```





```
nsclc cohort <- create analytic cohort(</pre>
  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",
```





```
nsclc cohort <- create analytic cohort(</pre>
  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
```





nsclc_cohort \$tbl_overall_ summary

Characteristic	N = 241 patients
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 coho	rt_ngs data frame
1	222 (92%)
2	18 (7.5%)
4	1 (0.4%)
¹ n (%)	

nsclc_cohort \$tbl_cohort

Characteristic	N = 241 Diagnoses		
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_squan	nous)		
Adenocarcinoma	241 (100%)		
¹ n (%)			

nsclc_cohort \$tbl_drugs

Characteristic	N = 241 Regimens
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

Characteristic	N = 262 Cancer Panel Tests
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 (%)	

Clinico-Genomic Data Processing Pipeline

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



Genomic Processing

{gnomeR}

create_gene_binary()

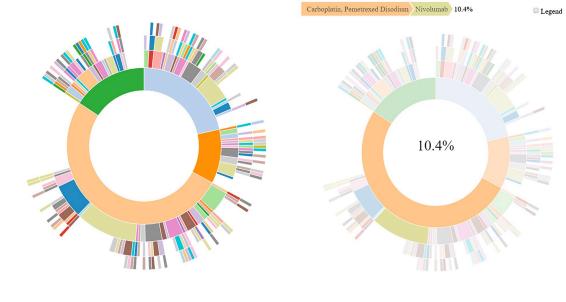
Processes data on mutation, CNA and fusions into analytic format

tbl_genomic()

Summarizes gene alterations across clinical variables of interest

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()

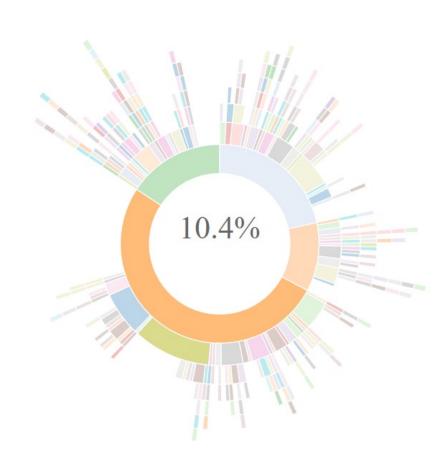
Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	 Name of object in global environment that was returned from pull_data_synapse()
data_cohort	The list returned from the create_analytic_cohort() function call	 Name of object in global environment that was returned 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





nsclc_sunburst\$ sunburst_plot



Clinico-Genomic Data Processing Pipeline

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



Genomic Processing

{gnomeR}

create_gene_binary()

Processes data on mutation, CNA and fusions into analytic format

tbl genomic()

Summarizes gene alterations across clinical variables of interest