

**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

## 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 **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.

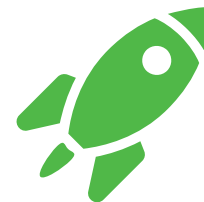
# Projects GENIE & GENIE BPC



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.

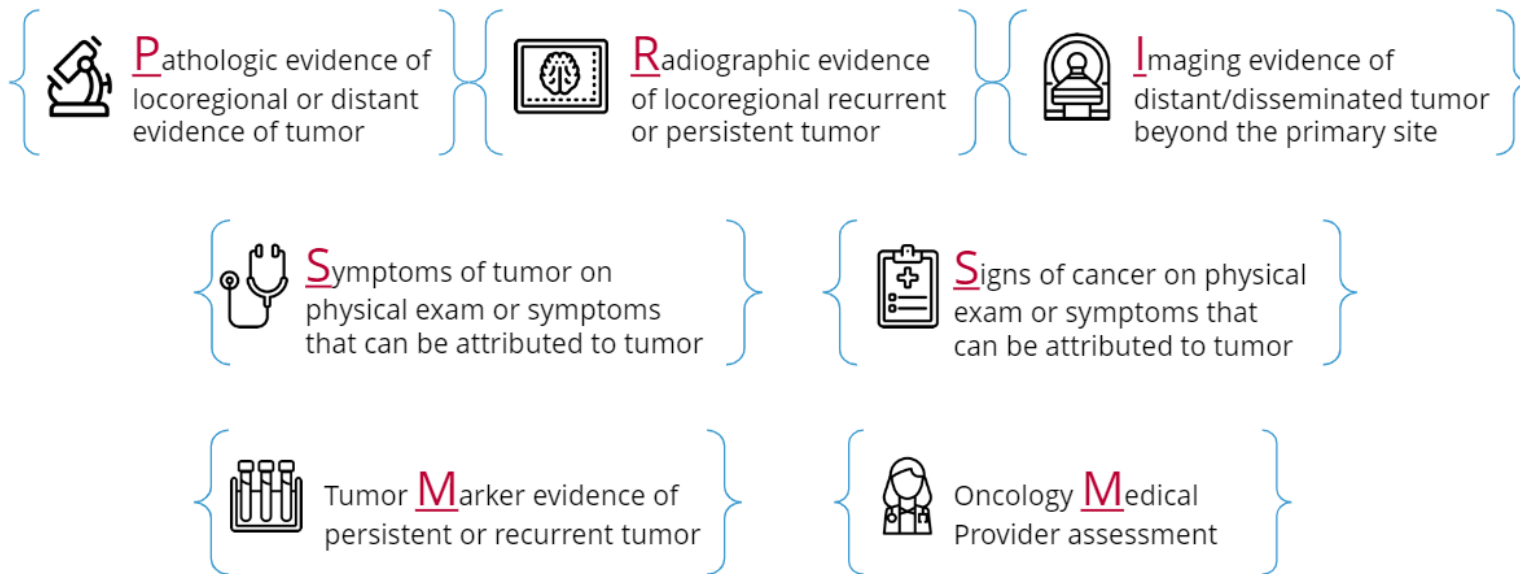


Phenomic data are curated using the PRISMM curation model to capture detailed information on cancer diagnosis, drug regimens, disease status from radiology reports, pathology reports and medical oncologist assessments, structured in several datasets with over 700 feature variables.



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.

## PRISSMM™: A Taxonomy for Defining Cancer Outcomes



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

- Data includes patients with  $\geq 1$  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

- 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

Files > Data Releases > NSCLC > 2.0-public > NSCLC\_2.0-public\_clinical\_data

 NSCLC\_2.0-public\_clinical\_data
 ★
📄
Download Options
⬇
Fo T

SynID [syn30358089](#) 
Items 8  
 Access  [View Terms](#)
Storage Location Synapse Storage

Name	Size
 cancer_level_dataset_index.csv	1.3 MB
 cancer_level_dataset_non_index.csv	323.3 KB
 cancer_panel_test_level_dataset.csv	562.2 KB
 imaging_level_dataset.csv	10.8 MB
 med_onc_note_level_dataset.csv	5.2 MB
 pathology_report_level_dataset.csv	3.9 MB
 patient_level_dataset.csv	512.2 KB
 regimen_cancer_level_dataset.csv	3.2 MB



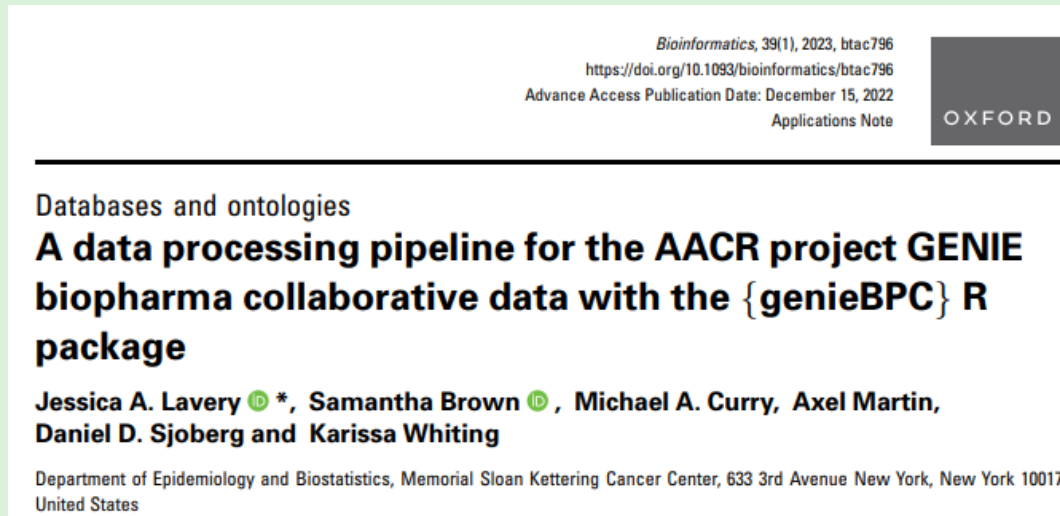
# {genieBPC} R Package

**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



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:

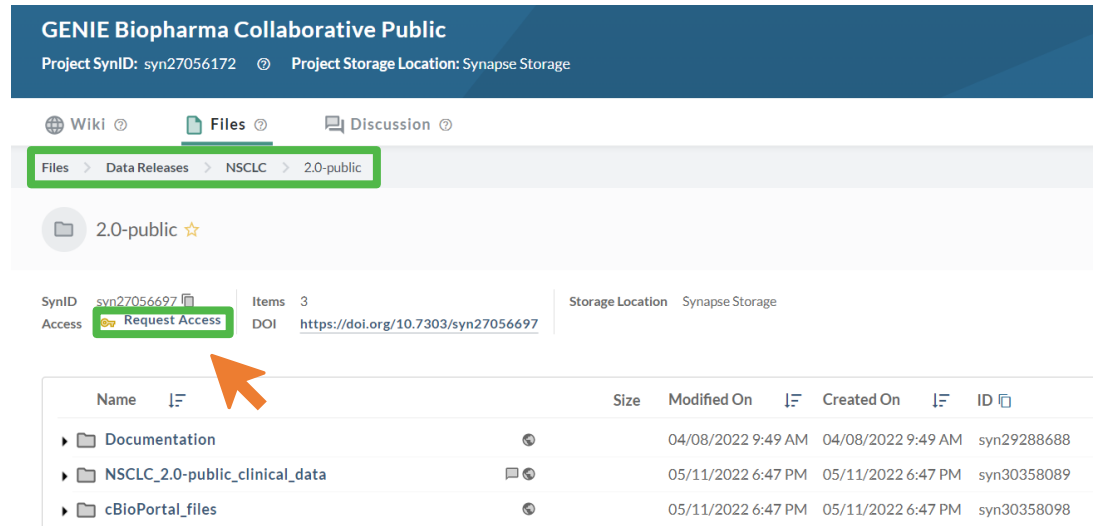
1. Register for a 'Synapse' account. Be sure to create a username and password. **Do NOT connect via your Google account.**

<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*



The screenshot displays the Synapse web interface for the 'GENIE Biopharma Collaborative Public' project. The project SynID is syn27056172 and the storage location is Synapse Storage. The breadcrumb navigation path is Files > Data Releases > NSCLC > 2.0-public, with the '2.0-public' folder highlighted. Below the breadcrumb, there is a folder icon and the text '2.0-public' with a star icon. The SynID is syn27056172, the number of items is 3, and the DOI is https://doi.org/10.7303/syn27056697. A green box highlights the 'Request Access' button, which is pointed to by an orange arrow. Below this, a table lists the files in the folder.

Name	Size	Modified On	Created On	ID
Documentation		04/08/2022 9:49 AM	04/08/2022 9:49 AM	syn29288688
NSCLC_2.0-public_clinical_data		05/11/2022 6:47 PM	05/11/2022 6:47 PM	syn30358089
cBioPortal_files		05/11/2022 6:47 PM	05/11/2022 6:47 PM	syn30358098

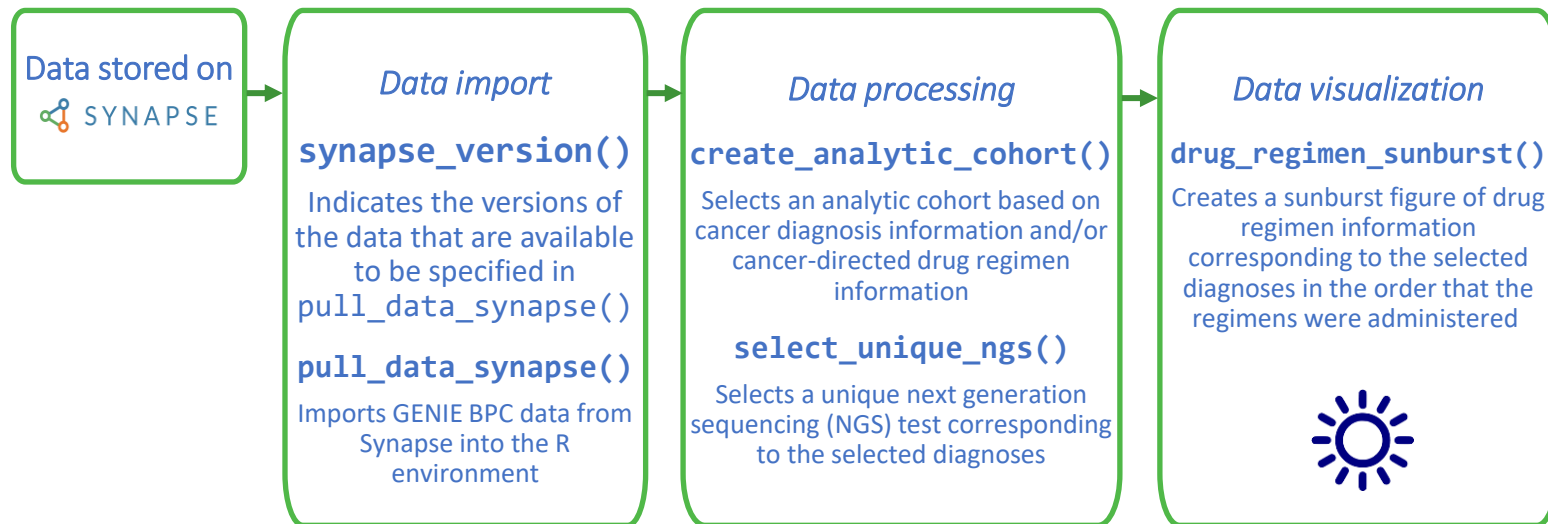
# Installation Instructions

## 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](https://github.com/GENIE-BPC/intro_to_genieBPC)
- Further R package details are available on the {genieBPC} [GitHub repo](#) & [website](#)
- {genieBPC} requires R version  $\geq 3.6$

# Clinico-Genomic Data Processing Pipeline

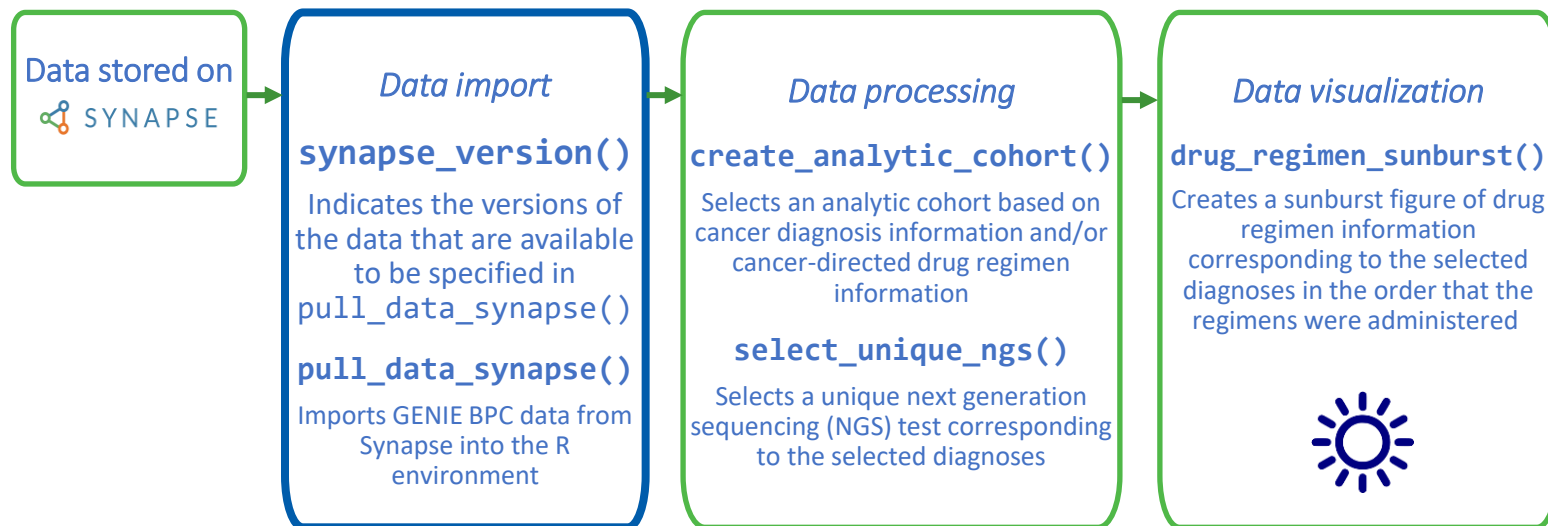


# Case Study

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](https://github.com/GENIE-BPC/intro_to_genieBPC)*

# Clinico-Genomic Data Processing Pipeline



# 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
<b>CRC</b>	<b>v2.0-public</b>	<b>October 2022</b>	<b>Most Recent Versions</b>
NSCLC	v2.2-consortium	February 2024	Most Recent Versions
<b>NSCLC</b>	<b>v2.0-public</b>	<b>May 2022</b>	<b>Most Recent Versions</b>
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	<ul style="list-style-type: none"> <li>• GENIE BPC Project cancer</li> <li>• Currently, NSCLC and CRC are the only two publicly available datasets</li> </ul>	<ul style="list-style-type: none"> <li>• NSCLC</li> <li>• CRC</li> <li>• BrCa</li> <li>• PANC</li> <li>• Prostate</li> <li>• BLADDER</li> </ul>
version	Version of the data (e.g v1.1-consortium, v2.0-public)	<ul style="list-style-type: none"> <li>• Values can be found in <code>synapse_version()</code></li> </ul>

## Demo: Run `pull_data_synapse()` for case study

```
library(genieBPC)

set_synapse_credentials()

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

**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: Run `pull_data_synapse()` for case study

```
library(genieBPC)

set_synapse_credentials()

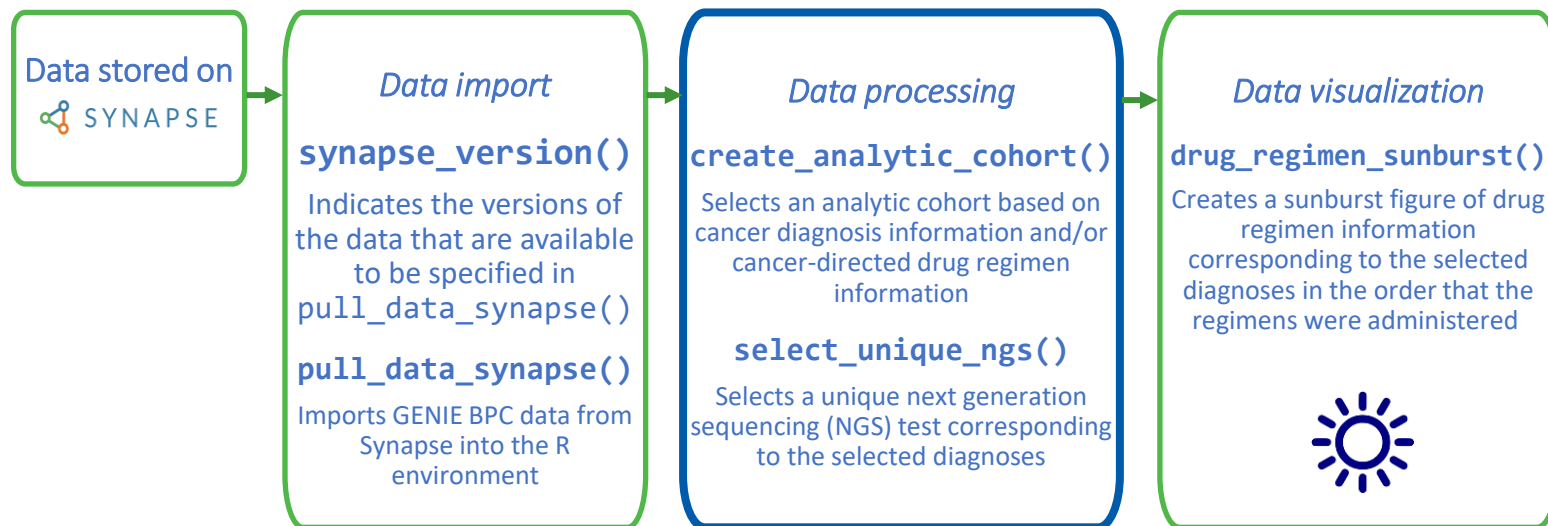
nscclc_synapse_data <-
  pull_data_synapse(
    cohort = "NSCLC",
    version = "v2.0-public")
```

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

**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



# Clinico-Genomic Data Processing Pipeline



```
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()	<ul style="list-style-type: none"><li>Name of object in global environment that was returned from pull_data_synapse()</li></ul>

# create\_analytic\_cohort()

Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	<ul style="list-style-type: none"><li>Name of object in global environment that was returned from pull_data_synapse()</li></ul>
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.	<ul style="list-style-type: none"><li>Numeric (1+)</li></ul>

# create\_analytic\_cohort()

Argument	Description	Acceptable Values
data_synapse	List returned from pull_data_synapse()	<ul style="list-style-type: none"><li>Name of object in global environment that was returned from pull_data_synapse()</li></ul>
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.	<ul style="list-style-type: none"><li>Numeric (1+)</li></ul>
institution	GENIE BPC participating institution. Default selection is all institutions. <i>Note that not all institutions curated data for all cancer sites.</i>	<ul style="list-style-type: none"><li>DFCI</li><li>MSK</li><li>UHN</li><li>VICC</li></ul>

# create\_analytic\_cohort()

Argument	Description	Acceptable Values
stage_dx	Stage at diagnosis. Default selection is all stages.	<ul style="list-style-type: none"><li>• Stage I</li><li>• Stage II</li><li>• Stage III</li><li>• Stage I-III NOS</li><li>• Stage IV</li></ul>
histology	<p>Cancer histology. Default selection is all histologies.</p> <p>For all cancer cohorts except for BrCa (breast cancer), this parameter corresponds to the variable 'ca_hist_adeno_squamous'.</p> <p>For BrCa, this parameter corresponds to the variable 'ca_hist_brca'</p>	<p>All cancer types except breast:</p> <ul style="list-style-type: none"><li>• Adenocarcinoma</li><li>• Squamous cell</li><li>• Sarcoma</li><li>• Small cell carcinoma</li><li>• Other histologies/mixed tumor</li></ul> <p>Breast cancer:</p> <ul style="list-style-type: none"><li>• Invasive lobular carcinoma</li><li>• Invasive ductal carcinoma</li><li>• Other histology</li></ul>

# 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_regimen_list 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.	<ul style="list-style-type: none"> <li>Exact</li> <li>Containing</li> </ul>

# Example: regimen\_drugs and regimen\_type

regimen_drugs	regimen_type	Example regimens returned
Carboplatin	Exact	<ul style="list-style-type: none"> <li>• Carboplatin</li> </ul>
Carboplatin	Containing	<ul style="list-style-type: none"> <li>• Carboplatin</li> <li>• Carboplatin, Cisplatin</li> <li>• Carboplatin, Paclitaxel</li> <li>• Carboplatin, Pemetrexed Disodium</li> <li>• etc.</li> </ul>

# 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.	<ul style="list-style-type: none"><li>Numeric (1+)</li></ul>
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")	<ul style="list-style-type: none"><li>Within cancer</li><li>Within regimen</li></ul>



## 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.	<ul style="list-style-type: none"><li>• TRUE</li><li>• FALSE</li></ul>

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

**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

**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

## Calling `nsc1c_cohort` returns a list of datasets:

**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

- `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`



```
nsc1c_cohort$  
tbl_overall_summary
```

## Overall Summary

Characteristic	N = 241 patients <sup>†</sup>
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%)
<sup>†</sup> n (%)	

```
nsc1c_cohort$  
tbl_cohort
```

## Cohort Summary

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

```
nsc1c_cohort$  
tbl_drugs
```

## Cancer-Directed Drugs Summary

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



```
nsc1c_cohort$  
tbl_ngs
```

# NGS Summary

Characteristic	N = 262 Cancer Panel Tests <sup>1</sup>
<b>Cohort (cohort)</b>	
NSCLC	262 (100%)
<b>Institution (institution)</b>	
DFCI	99 (38%)
MSK	126 (48%)
VICC	37 (14%)
<b>OncoTree code (cpt_oncotree_code)</b>	
LCLC	1 (0.4%)
LUAD	253 (97%)
LUAS	1 (0.4%)
LUSC	1 (0.4%)
NSCLC	4 (1.5%)
NSCLCPD	2 (0.8%)
<b>Sequence assay ID (cpt_seq_assay_id)</b>	
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%)
<sup>1</sup> n (%)	

# select\_unique\_ngs()



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	<ul style="list-style-type: none"><li>Name of NGS object in global environment that was returned from create_analytic_cohort()</li></ul>
oncotree_code	Character vector specifying which sample OncoTree codes to prioritize.	<ul style="list-style-type: none"><li>See 'cpt_oncotree_code' column of data_cohort.</li></ul>
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.	<ul style="list-style-type: none"><li>Primary</li><li>Local</li><li>Metastasis</li></ul>
min_max_time	Character specifying if the first or last genomic sample recorded should be kept.	<ul style="list-style-type: none"><li>min (refers to earliest sample)</li><li>max (refers to latest sample)</li></ul>

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

**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(nscclc_cohort$cohort_ngs)
```

```
[1] 262
```

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

```
nrow(nscclc_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

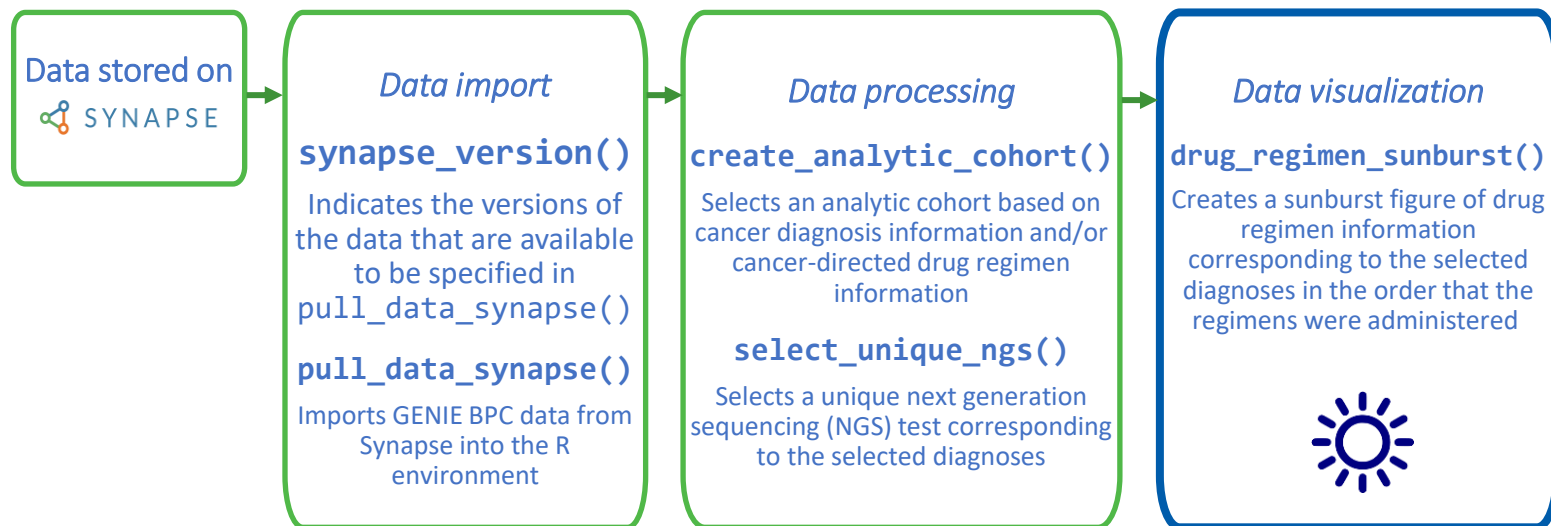
## Visualization

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



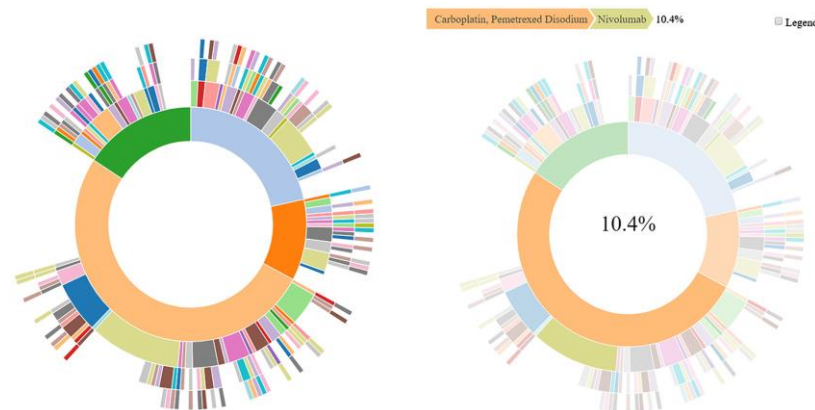
See <https://mskcc-epi-bio.github.io/gnomeR/> for more details.

# Clinico-Genomic Data Processing Pipeline



## 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()	<ul style="list-style-type: none"><li>Name of object in global environment that was returned from pull_data_synapse()</li></ul>
data_cohort	The list returned from the create_analytic_cohort() function call	<ul style="list-style-type: none"><li>Name of object in global environment that was returned from create_analytic_cohort()</li></ul>
max_n_regimens	The maximum number of regimens displayed in the sunburst plot	<ul style="list-style-type: none"><li>Integer &gt;0</li></ul>



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

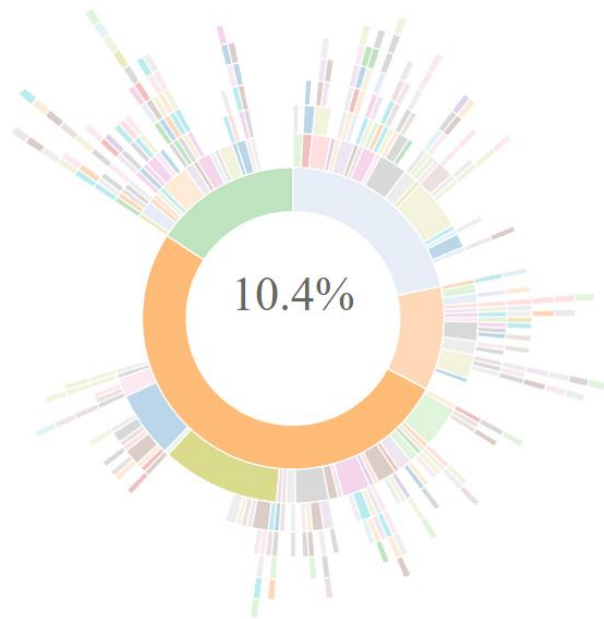
**Case Study:** Create a cohort of patients who were diagnosed with **Stage IV adenocarcinoma NSCLC** and received **Carboplatin and Pemetrexed +/- Bevacizumab or Cisplatin and Pemetrexed +/- Bevacizumab** as their **first** cancer-directed drug regimen after diagnosis

```
nsc1c_sunburst <- drug_regimen_sunburst(  
  data_synapse = nsc1c_synapse_data$NSCLC_v2.0,  
  data_cohort = nsc1c_cohort)
```



```
nsc1c_sunburst$  
sunburst_plot
```

Carboplatin, Pemetrexed Disodium Nivolumab 10.4%



## 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

- 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 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 Mastrogiamomo*  
*Ino de Bruijn*

## **Statistical Coordinating Center**

*Kathy Panageas*  
*Jessica Lavery*  
*Samantha Brown*  
*Hannah Fuchs*