G00x - Generalizable Germline-Targeting Clinical Trial Pipeline

- C de f r generating main text figures 2 thr ugh 7.
- C de f r generating the spreadsheet Data S10. “S urce data”.

Sec nd, the rep sit ry presents the G00x pipeline designed t facilitate the analysis f germline-targeting vaccine clinical trials. It is an all-in- ne pipeline and analysis platf rm that parses, validates, calculates B cell frequencies, runs cellranger and c mbines all analyses int a pl ttable dataframe. The pipeline supp rts datasets f r b th G002 and G003 trials (related t germline targeting t elicit VRC01-class antib dies against HIV) and ensures that all data is pr cessed and validated t maintain the integrity f the clinical trial results. The pipeline is readily m difiable t enable handling f ther types f germline-targeting trials, with ther types f analyses, and can als be m dified t enable data st rage/data fl w pathways different fr m that used in G002 and G003.

Data Access

If y u d n't want t run the pipeline, y u can access the imp rtant data via the f ll wing links.

- The ann tated, filtered and paired antib dy sequences:
  - G002 final_df.feather
  - G003 final_df.feather
- Merged summary file with all frequencies rep rted in this study:
  - G002 fl w_and_sequencing_l ng_names.csv
  - G003 fl w_and_sequencing_l ng_names.csv

G002

T run testing, y u will need a `g002` direct ry in the r t f the pr ject. Y u can get this with the sync c mmand:

```
# this will sync the entire contents of the files into g002 directory
# Beware, this file is 1.5TB
aws sync --delete s3 s3://iavig002public/g002/ ./g002'
```

H wever, we have als setup s y u d n't need t d wnl ad the entire direct ry. Y u can run the f ll wing t get a subset f the data.

```
# get the sorting directory
aws s3 cp --recursive s3://iavig002public/g002/G002/sorting ./g002/G002/sorting

# get the sequencing directory excluding large bcl and fastq files
aws s3 cp --recursive s3://iavig002public/g002/G002/sequencing ./g002/G002/sequencing --exclude *working_directory/*
--exclude *.fastq.gz --exclude *.tif --exclude *.cbcl --exclude *.imf1 --exclude *.filter --exclude *.bin --exclude
*Logs/* --exclude *_stdout --exclude *_stderr --exclude *Autofocus/* --exclude *Intensities/*
```

G003

T run testing, y u will need a `g003` direct ry in the r t f the pr ject. Y u can get this with the sync c mmand

```
# this will sync the entire contents of the files into g003 directory
# Beware, this file is 1.5TB
aws sync --delete s3 s3://iavig003public/g003/ ./g003'
```

H wever, we have als setup s y u d n't need t d wnl ad the entire direct ry. Y u can run the f ll wing t get a subset f the data.

```
# get the sorting directory
aws s3 cp --recursive s3://iavig003public/g003/G003/sorting ./g003/G003/sorting

# get the sequencing directory excluding large bcl and fastq files
aws s3 cp --recursive s3://iavig003public/g003/G003/sequencing ./g003/G003/sequencing --exclude *working_directory/*
--exclude *.fastq.gz --exclude *.tif --exclude *.cbcl --exclude *.imf1 --exclude *.filter --exclude *.bin --exclude
*Logs/* --exclude *_stdout --exclude *_stderr --exclude *Autofocus/* --exclude *Intensities/*
```

Pipeline

Installati n pre-requisites

While n t necessary, we highly rec mmend using the c nda pen-s urce package and envir nment manager. F r the purp ses f this rep sit ry, nly a minimal installer f r anac nda is necessary (Minic nda).

Minic nda c mmand line installers

Installati n

This installati n assumes that `git` and `conda` are in y ur path.

```
# clone the repository
git clone https://github.com/SchiefLab/G00x.git

# change directory
cd G00x

# this will create a conda environment called g00x and install the package
./install.sh
```

Validati n

The m st imp rtant part f this clinical trial ther than it w rking, is that everything is validated. That means that the data is validated, the c de is validated, and the results are validated. This is a very imp rtant part f the pr cess and sh uld n t be skipped.

FACS/S rting validati n

G002

These are instructi ns n h w t validate the file structure c ntaining the FACS/S rting data bef re upl ading t b x.

```
#Run the validator for flow from the command line
g00x g002 validate flow my_path/to/box/G002/


# If you used the AWS sync command above, you can use the command:
g00x g002 validate flow ./g002/G002/sorting/G002
```

Y ur f lder structure n b x sh uld l k like this.

```
# if you used the above commands to sync the data. The folder structure will look like this

g002/G002/sorting
└── G002
    ├── Prescreens
    │   ├── Prescreen_RunDate220825_UploadDate221021
    │   ├── Prescreen_RunDate220826_UploadDate221021
    ....
    └── Sorts
        ├── Sort_RunDate220927_UploadDate221013
        ├── Sort_RunDate220928_UploadDate221014
        ├── Sort_RunDate220929_UploadDate221014
        ├── Sort_RunDate220930_UploadDate221014
```

G003

```
#Run the validator for flow from the command line
g00x g003 validate flow my_path/to/G003/


# If you used the AWS sync command above, you can use th e
g00x g003 validate flow ./g003/G003/sorting/G003
```

F r G003, all the data is available in S3 bucket. The structure sh uld l k as bel w:

```
g003/G003/sorting
└── G003
    ├── Prescreens
    │   ├── Sort_RunDate230807_UploadDate230807
    │   ├── Sort_RunDate230808_UploadDate230808
    ....
    └── Sorts
        ├── Sort_RunDate230807_UploadDate230807
        ├── Sort_RunDate230808_UploadDate230808
        ├── Sort_RunDate230809_UploadDate230809
        ├── Sort_RunDate231003_UploadDate231027
```

Sequencing files validati n

G002

In rder t validate the sequencing files, it must be merged with the fl w data. Thus, y u will need b th Gl bus and B x access. Validati n can be acc mplished by validating the merge c mmand

```
g00x g002 sequencing –f /path/to/flow –s /path/to/sequencing –o output_file


# if you have the AWS structure
g00x g002 merge –s ./g002/G002/sequencing/G002 –f ./g002/G002/sorting/G002 –o my_merged_file
```

The sequencing f lder structure sh uld be as f ll ws:

```
G002
├── run0002
│   ├── 221006_VH00497_31_AAAVKCLHV
│   └── sample_manifest.csv
├── run0003
│   ├── 221019_VH00497_32_AAANGGVM5
│   └── sample_manifest.csv
└── run0004
    ├── 221101_VL00414_3_AACFYLCM5
    ├── 221103_VL00414_4_AAATJGYM5
    └── sample_manifest.csv
```

G003

In G003, all the data are st red in the S3 bucket; thus, y u nly need access.

```
g00x g003 sequencing –f /path/to/flow –s /path/to/sequencing –o output_file


# if you have the AWS structure
g00x g003 merge –s ./g003/G003/sequencing/G003 –f ./g003/G003/sorting/G003 –o my_merged_file
```

```
G003
├── run0001
│   ├── 230818_NB552490_0059_AHL7GFBGXM
│   └── sequencing_manifest.csv
├── run0002
│   ├── 230821_NB552490_0060_AHL7GGBGXM
│   └── sequencing_manifest.csv
├── run0003
│   ├── 231108_NB552490_0065_AHL7CCBGXM
│   └── sequencing_manifest.csv
```

Using the ab ve c mmand, y u will generate a file called `my_merged_file.csv` which will have the f ll wing fields.

```
ptid
group
weeks
visit_id
probe_set
sample_type
run_date
sort_pool
hashtag
run_dir_path
pool_number
sorted_date
vdj_sequencing_replicate
cso_sequencing_replicate
vdj_lirary_replicate
cso_library_replicate
bio_replicate
vdj_index
```

fi lds	d scription
value	what is the value f the gate r frequency
hashtag	if this sample has been s rted, what is the hashtag
s rt_p l	the s rting p l, P01 - P10

BCR sequence analysis

G002

```
g00x g002 pipeline demultiplex -f ./g002/G002/sorting/G002/ -s ./g002/G002/sequencing/G002/ -o
./g002/G002/output/demultiplex

g00x g002 pipeline vdj -d ./g002/G002/output/demultiplex.feather -o ./g002/G002/output/vdj

g00x g002 pipeline cso -d ./g002/G002/output/demultiplex.feather -o ./g002/G002/output/cso

# Merge VDJ and CSO dataframes, run the output through SADIE for
g00x g002 pipeline airr -k 3 -c ./g002/G002/output/cso.feather -v ./g002/G002/output/vdj.feather -o
./g002/G002/output/final_df

# Output merged.feather not used to generate figures, but is a sanity check.
g00x g002 validate merge -s ./g002/G002/sequencing/G002/ -f ./g002/G002/sorting/G002 -o ./g002/G002/output/merged

# Output flow_and_sequencing.feather used to generate figures; will contain counts, frequencies, and metadata from
flow and sequencing data.
g00x g002 analysis report -s ./g002/G002/output/final_df.feather -f ./g002/G002/output/flow_output.feather -o
./g002/G002/output/flow_and_sequencing
```

G003

```
g00x g003 pipeline demultiplex -f ./g003/G003/sorting/G003/ -s ./g003/G003/sequencing/G003/ -o
./g003/G003/output/demultiplex

g00x g003 pipeline vdj -d ./g003/G003/output/demultiplex.feather -o ./g003/G003/output/vdj

g00x g003 pipeline cso -d ./g003/G003/output/demultiplex.feather -o ./g003/G003/output/cso

# Merge VDJ and CSO dataframes, run the output through SADIE for
g00x g003 pipeline airr -k 3 -c ./g003/G003/output/cso.feather -v ./g003/G003/output/vdj.feather -o
./g003/G003/output/final_df

# Output merged.feather not used to generate figures, but is a sanity check.
g00x g003 validate merge -s ./g003/G003/sequencing/G003/ -f ./g003/G003/sorting/G003 -o ./g003/G003/output/merged

# Output flow_and_sequencing.feather used to generate figures; will contain counts, frequencies, and metadata from
flow and sequencing data.
g00x g003 analysis report -s ./g003/G003/output/final_df.feather -f ./g003/G003/output/flow_output.feather -o
./g003/G003/output/flow_and_sequencing
```

Devel pment

Testing

If y u'd like t help devel p. Please use pre-c mmit bef re c mmiting any files.

```
# runs pre-commit for syntax,formatting and static type checking
poetry run pre-commit run --all-files
```

N w y u can run the tests via p etry

```
poetry run pytest -sv --log-cli-level DEBUG tests
```

Figures and Tables

Main Figures

Main figures as they are displayed in the G002 and G003 paper. Figure 1 is n t included as it is manually generated.

```
g00x plots fig2
g00x plots fig3
g00x plots fig4
g00x plots fig5
g00x plots fig6
g00x plots fig7
```

Supplementary Figures

Supplementary figures are in rder.

```
g00x plots suppfigures --all #TODO: Update later
```

Supplementary C mparis n Tables

```
g00x plots comparison-tables --all
g00x plots tables --all
```

Issues

Please submit any issues t the issues page and we are happy t help.

License

L PL