RTCGAToolbox

Get ready

1. Install R: https://cran.rstudio.com/
2. Install RStudio: https://posit.co/download/rstudio-desktop/
3. Install RTCGAToolbox:

```
if (!requireNamespace("BiocManager"))
    install.packages("BiocManager")
BiocManager::install("RTCGAToolbox")
```

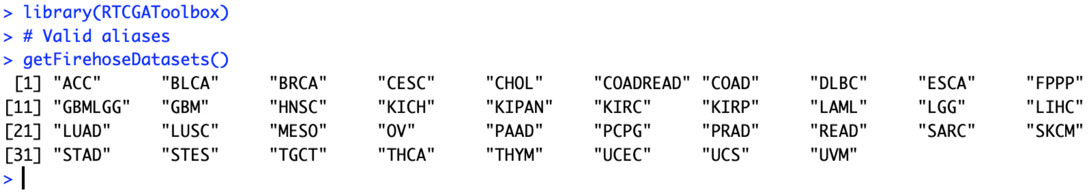
Exploration

Data Client

1. **Check valid dataset aliases, stddata run dates and analyze run dates:**

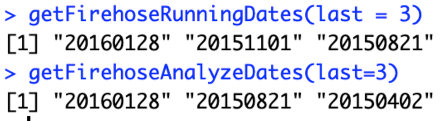
getFirehoseDatasets() – valid dataset aliases
getFirehoseRunningDates() – stddata run dates (???)
getFirehoseAnalyzeDates() – analyze run dates

```
library(RTCGAToolbox)
# Valid aliases
getFirehoseDatasets()
```



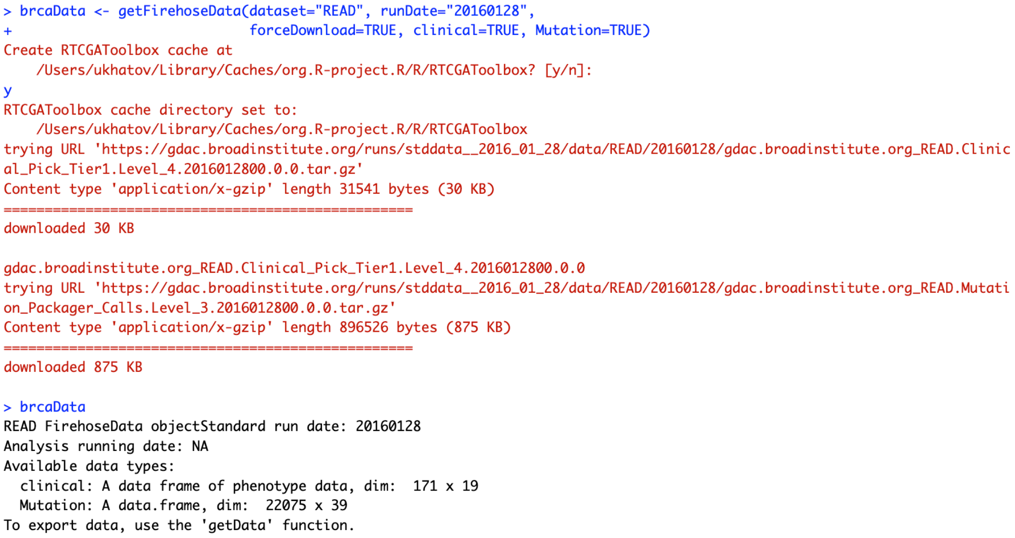
```
> library(RTCGAToolbox)
> # Valid aliases
> getFirehoseDatasets()
 [1] "ACC"      "BLCA"     "BRCA"     "CESC"     "CHOL"     "COADREAD" "COAD"     "DLBC"     "ESCA"     "FPPP"
[11] "GBMLGG"   "GBM"      "HNSC"     "KICH"     "KIPAN"    "KIRC"     "KIRP"     "LAML"     "LGG"      "LIHC"
[21] "LUAD"     "LUSC"     "MESO"     "OV"       "PAAD"     "PCPG"     "PRAD"     "READ"     "SARC"     "SKCM"
[31] "STAD"     "STES"     "TGCT"     "THCA"     "THYM"     "UCEC"     "UCS"      "UVM"
> |
```

```
# Valid stddata runs
getFirehoseRunningDates(last=3)
# Valid analysis running dates (will return 3 recent date)
getFirehoseAnalyzeDates(last=3)
```



```
> getFirehoseRunningDates(last = 3)
[1] "20160128" "20151101" "20150821"
> getFirehoseAnalyzeDates(last=3)
[1] "20160128" "20150821" "20150402"
```

```
# READ mutation data and clinical data
brcaData <- getFirehoseData(dataset="READ", runDate="20160128",
    forceDownload=TRUE, clinical=TRUE, Mutation=TRUE)
brcaData
```

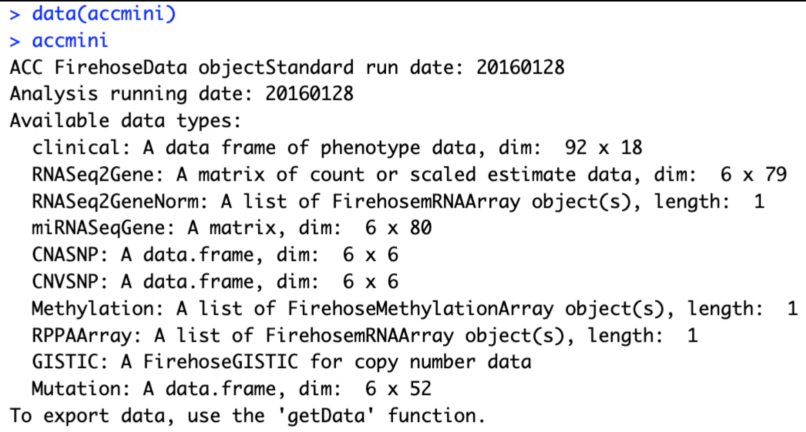


```
> brcaData <- getFirehoseData(dataset="READ", runDate="20160128",
+                             forceDownload=TRUE, clinical=TRUE, Mutation=TRUE)
Create RTCGAToolbox cache at
    /Users/ukhatov/Library/Caches/org.R-project.R/R/RTCGAToolbox? [y/n]:
y
RTCGAToolbox cache directory set to:
    /Users/ukhatov/Library/Caches/org.R-project.R/R/RTCGAToolbox
trying URL 'https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/READ/20160128/gdac.broadinstitute.org_READ.Clinic
al_Pick_Tier1.Level_4.2016012800.0.0.tar.gz'
Content type 'application/x-gzip' length 31541 bytes (30 KB)
==================================================
downloaded 30 KB

gdac.broadinstitute.org_READ.Clinical_Pick_Tier1.Level_4.2016012800.0.0
trying URL 'https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/READ/20160128/gdac.broadinstitute.org_READ.Mutati
on_Packager_Calls.Level_3.2016012800.0.0.tar.gz'
Content type 'application/x-gzip' length 896526 bytes (875 KB)
==================================================
downloaded 875 KB

> brcaData
READ FirehoseData objectStandard run date: 20160128
Analysis running date: NA
Available data types:
  clinical: A data frame of phenotype data, dim:  171 x 19
  Mutation: A data.frame, dim:  22075 x 39
To export data, use the 'getData' function.
```

2. Example Dataset Exploration

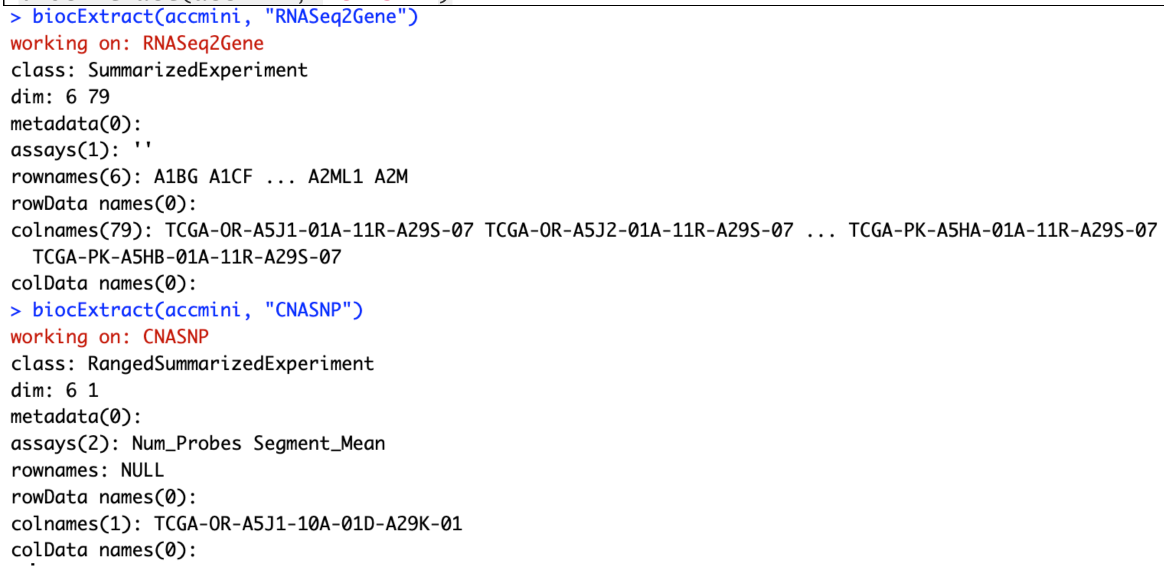
Accmini – 'ACC' (Adrenocortical carcinoma) that contains only the top 6 rows for each dataset and a full clinical dataset.

```
data(accmini)
accmini
```



```
> data(accmini)
> accmini
ACC FirehoseData objectStandard run date: 20160128
Analysis running date: 20160128
Available data types:
  clinical: A data frame of phenotype data, dim:  92 x 18
  RNASeq2Gene: A matrix of count or scaled estimate data, dim:  6 x 79
  RNASeq2GeneNorm: A list of FirehosemRNAArray object(s), length:  1
  miRNASeqGene: A matrix, dim:  6 x 80
  CNASNP: A data.frame, dim:  6 x 6
  CNVSNP: A data.frame, dim:  6 x 6
  Methylation: A list of FirehoseMethylationArray object(s), length:  1
  RPPAArray: A list of FirehosemRNAArray object(s), length:  1
  GISTIC: A FirehoseGISTIC for copy number data
  Mutation: A data.frame, dim:  6 x 52
To export data, use the 'getData' function.
```

```
biocExtract(accmini, "RNASeq2Gene")
biocExtract(accmini, "CNASNP")
```



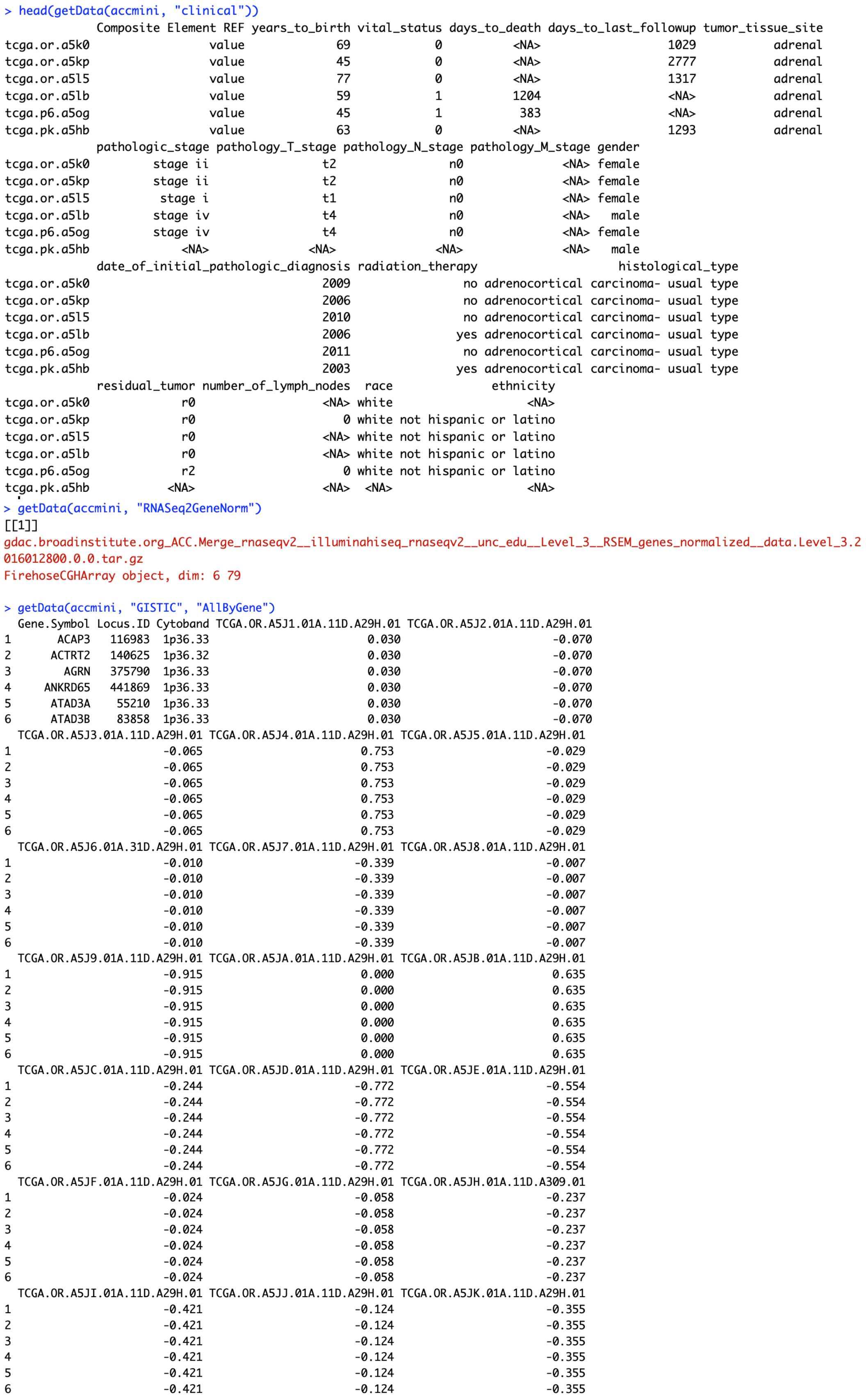
```
> biocExtract(accmini, "RNASeq2Gene")
working on: RNASeq2Gene
class: SummarizedExperiment
dim: 6 79
metadata(0):
assays(1): ''
rownames(6): A1BG A1CF ... A2ML1 A2M
rowData names(0):
colnames(79): TCGA-OR-A5J1-01A-11R-A29S-07 TCGA-OR-A5J2-01A-11R-A29S-07 ... TCGA-PK-A5HA-01A-11R-A29S-07
  TCGA-PK-A5HB-01A-11R-A29S-07
colData names(0):
> biocExtract(accmini, "CNASNP")
working on: CNASNP
class: RangedSummarizedExperiment
dim: 6 1
metadata(0):
assays(2): Num_Probes Segment_Mean
rownames: NULL
rowData names(0):
colnames(1): TCGA-OR-A5J1-10A-01D-A29K-01
colData names(0):
```

Following logic keys are provided for different data types. By default client only download clinical data.

- RNAseqGene
- clinical
- RNASeqGene
- RNASeq2Gene
- RNASeq2GeneNorm
- miRNASeqGene
- CNASNP
- CNVSNP
- CNASeq
- CNACGH
- Methylation
- Mutation
- mRNAArray
- miRNAArray
- RPPAArray

3. Raw Data

```
head(getData(accmini, "clinical"))
getData(accmini, "RNASeq2GeneNorm")
getData(accmini, "GISTIC", "AllByGene")
```



```
> head(getData(accmini, "clinical"))
             Composite Element REF years_to_birth vital_status days_to_death days_to_last_followup tumor_tissue_site
tcga.or.a5k0                 value             69            0          <NA>                  1029           adrenal
tcga.or.a5kp                 value             45            0          <NA>                  2777           adrenal
tcga.or.a5l5                 value             77            0          <NA>                  1317           adrenal
tcga.or.a5lb                 value             59            1          1204                  <NA>           adrenal
tcga.p6.a5og                 value             45            1           383                  <NA>           adrenal
tcga.pk.a5hb                 value             63            0          <NA>                  1293           adrenal
             pathologic_stage pathology_T_stage pathology_N_stage pathology_M_stage gender
tcga.or.a5k0         stage ii                t2                n0              <NA> female
tcga.or.a5kp         stage ii                t2                n0              <NA> female
tcga.or.a5l5          stage i                t1                n0              <NA> female
tcga.or.a5lb         stage iv                t4                n0              <NA>   male
tcga.p6.a5og         stage iv                t4                n0              <NA> female
tcga.pk.a5hb             <NA>              <NA>              <NA>              <NA>   male
             date_of_initial_pathologic_diagnosis radiation_therapy                    histological_type
tcga.or.a5k0                                 2009                no adrenocortical carcinoma- usual type
tcga.or.a5kp                                 2006                no adrenocortical carcinoma- usual type
tcga.or.a5l5                                 2010                no adrenocortical carcinoma- usual type
tcga.or.a5lb                                 2006               yes adrenocortical carcinoma- usual type
tcga.p6.a5og                                 2011                no adrenocortical carcinoma- usual type
tcga.pk.a5hb                                 2003               yes adrenocortical carcinoma- usual type
             residual_tumor number_of_lymph_nodes  race              ethnicity
tcga.or.a5k0             r0                  <NA> white                   <NA>
tcga.or.a5kp             r0                     0 white not hispanic or latino
tcga.or.a5l5             r0                  <NA> white not hispanic or latino
tcga.or.a5lb             r0                  <NA> white not hispanic or latino
tcga.p6.a5og             r2                     0 white not hispanic or latino
tcga.pk.a5hb           <NA>                  <NA>  <NA>                   <NA>
> getData(accmini, "RNASeq2GeneNorm")
[[1]]
gdac.broadinstitute.org_ACC.Merge_rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.Level_3.2
016012800.0.0.tar.gz
FirehoseCGHArray object, dim: 6 79

> getData(accmini, "GISTIC", "AllByGene")
  Gene.Symbol Locus.ID Cytoband TCGA.OR.A5J1.01A.11D.A29H.01 TCGA.OR.A5J2.01A.11D.A29H.01
1       ACAP3   116983  1p36.33                        0.030                       -0.070
2      ACTRT2   140625  1p36.32                        0.030                       -0.070
3        AGRN   375790  1p36.33                        0.030                       -0.070
4     ANKRD65   441869  1p36.33                        0.030                       -0.070
5      ATAD3A    55210  1p36.33                        0.030                       -0.070
6      ATAD3B    83858  1p36.33                        0.030                       -0.070
  TCGA.OR.A5J3.01A.11D.A29H.01 TCGA.OR.A5J4.01A.11D.A29H.01 TCGA.OR.A5J5.01A.11D.A29H.01
1                       -0.065                        0.753                       -0.029
2                       -0.065                        0.753                       -0.029
3                       -0.065                        0.753                       -0.029
4                       -0.065                        0.753                       -0.029
5                       -0.065                        0.753                       -0.029
6                       -0.065                        0.753                       -0.029
  TCGA.OR.A5J6.01A.31D.A29H.01 TCGA.OR.A5J7.01A.11D.A29H.01 TCGA.OR.A5J8.01A.11D.A29H.01
1                       -0.010                       -0.339                       -0.007
2                       -0.010                       -0.339                       -0.007
3                       -0.010                       -0.339                       -0.007
4                       -0.010                       -0.339                       -0.007
5                       -0.010                       -0.339                       -0.007
6                       -0.010                       -0.339                       -0.007
  TCGA.OR.A5J9.01A.11D.A29H.01 TCGA.OR.A5JA.01A.11D.A29H.01 TCGA.OR.A5JB.01A.11D.A29H.01
1                       -0.915                        0.000                        0.635
2                       -0.915                        0.000                        0.635
3                       -0.915                        0.000                        0.635
4                       -0.915                        0.000                        0.635
5                       -0.915                        0.000                        0.635
6                       -0.915                        0.000                        0.635
  TCGA.OR.A5JC.01A.11D.A29H.01 TCGA.OR.A5JD.01A.11D.A29H.01 TCGA.OR.A5JE.01A.11D.A29H.01
1                       -0.244                       -0.772                       -0.554
2                       -0.244                       -0.772                       -0.554
3                       -0.244                       -0.772                       -0.554
4                       -0.244                       -0.772                       -0.554
5                       -0.244                       -0.772                       -0.554
6                       -0.244                       -0.772                       -0.554
  TCGA.OR.A5JF.01A.11D.A29H.01 TCGA.OR.A5JG.01A.11D.A29H.01 TCGA.OR.A5JH.01A.11D.A309.01
1                       -0.024                       -0.058                       -0.237
2                       -0.024                       -0.058                       -0.237
3                       -0.024                       -0.058                       -0.237
4                       -0.024                       -0.058                       -0.237
5                       -0.024                       -0.058                       -0.237
6                       -0.024                       -0.058                       -0.237
  TCGA.OR.A5JI.01A.11D.A29H.01 TCGA.OR.A5JJ.01A.11D.A29H.01 TCGA.OR.A5JK.01A.11D.A29H.01
1                       -0.421                       -0.124                       -0.355
2                       -0.421                       -0.124                       -0.355
3                       -0.421                       -0.124                       -0.355
4                       -0.421                       -0.124                       -0.355
5                       -0.421                       -0.124                       -0.355
6                       -0.421                       -0.124                       -0.355
```

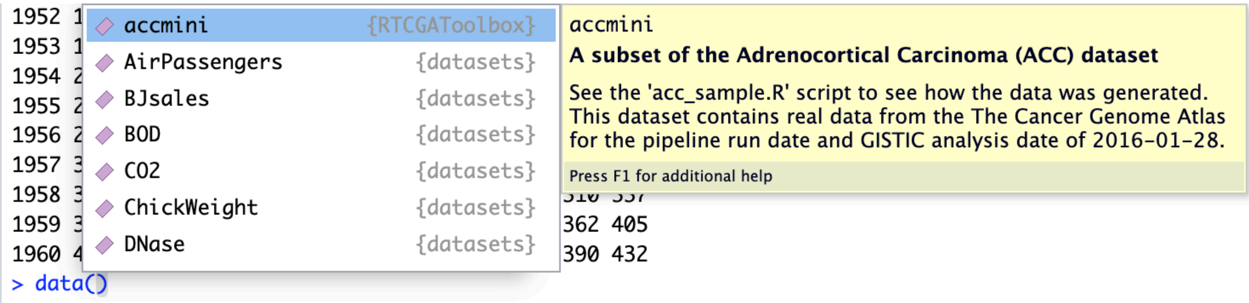
4. Session information

```
sessionInfo()
```

```
> sessionInfo()
R version 4.2.2 (2022-10-31)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Monterey 12.2.1

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] RTCGAToolbox_2.28.0

loaded via a namespace (and not attached):
 [1] MatrixGenerics_1.10.0      Biobase_2.58.0             httr_1.4.5                 bit64_4.0.5
 [5] jsonlite_1.8.4             splines_4.2.2              TCGAutils_1.18.0           BiocManager_1.30.20
 [9] stats4_4.2.2               BiocFileCache_2.6.1        blob_1.2.3                 Rsamtools_2.14.0
[13] GenomeInfoDbData_1.2.9     yaml_2.3.7                 progress_1.2.2             pillar_1.8.1
[17] RSQLite_2.3.0              lattice_0.20-45            glue_1.6.2                 limma_3.54.2
[21] digest_0.6.31              GenomicRanges_1.50.2       XVector_0.38.0             rvest_1.0.3
[25] Matrix_1.5-3               XML_3.99-0.13              pkgconfig_2.0.3            biomaRt_2.54.0
[29] zlibbioc_1.44.0            RCircos_1.2.2              MultiAssayExperiment_1.24.0 BiocParallel_1.32.5
[33] tzdb_0.3.0                 tibble_3.1.8               KEGGREST_1.38.0            generics_0.1.3
[37] IRanges_2.32.0             ellipsis_0.3.2             cachem_1.0.7               SummarizedExperiment_1.28.0
[41] GenomicFeatures_1.50.4     BiocGenerics_0.44.0        cli_3.6.0                  survival_3.5-3
[45] RJSONIO_1.3-1.8            magrittr_2.0.3             crayon_1.5.2               memoise_2.0.1
[49] fansi_1.0.4                xml2_1.3.3                 tools_4.2.2                data.table_1.14.8
[53] prettyunits_1.1.1          hms_1.1.2                  BiocIO_1.8.0               lifecycle_1.0.3
[57] matrixStats_0.63.0         stringr_1.5.0              S4Vectors_0.36.2           DelayedArray_0.24.0
[61] AnnotationDbi_1.60.0       Biostrings_2.66.0          compiler_4.2.2             GenomeInfoDb_1.34.9
[65] rlang_1.0.6                grid_4.2.2                 GenomicDataCommons_1.22.1  RCurl_1.98-1.10
[69] rjson_0.2.21               rappdirs_0.3.3             bitops_1.0-7               codetools_0.2-19
[73] restfulr_0.0.15            DBI_1.1.3                  curl_5.0.0                 R6_2.5.1
[77] GenomicAlignments_1.34.0   rtracklayer_1.58.0         dplyr_1.1.0                fastmap_1.1.1
[81] bit_4.0.5                  utf8_1.2.3                 filelock_1.0.2             readr_2.1.4
[85] stringi_1.7.12             parallel_4.2.2             RaggedExperiment_1.22.0    Rcpp_1.0.10
[89] vctrs_0.5.2                png_0.1-8                  dbplyr_2.3.1               tidyselect_1.2.0
```

Questions:

1) What is “stddata run dates”?
2) “accmini (ACC)”. How to find this dataset without guide? Which datasets are also available? How to find dataset "BLCA", "BRCA", "CESC", etc…



```
1952 1
1953 1
1954 2
1955 2
1956 2
1957 3
1958 3
1959 3     362 405
1960 4     390 432
> data()
```

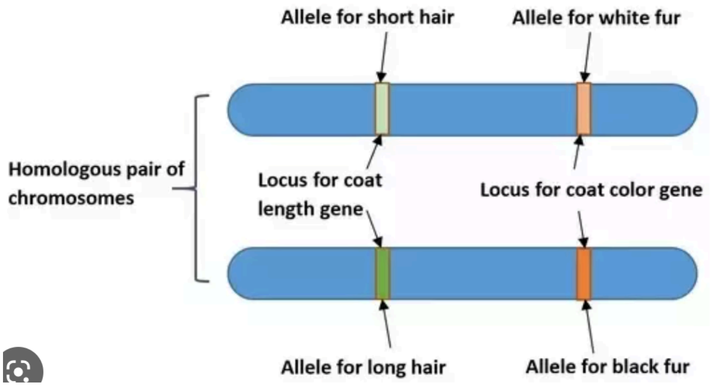
accmini

A subset of the Adrenocortical Carcinoma (ACC) dataset

See the 'acc_sample.R' script to see how the data was generated. This dataset contains real data from the The Cancer Genome Atlas for the pipeline run date and GISTIC analysis date of 2016-01-28.

```
> data("AirPassengers")
> AirPassengers
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1949 112 118 132 129 121 135 148 148 136 119 104 118
1950 115 126 141 135 125 149 170 170 158 133 114 140
1951 145 150 178 163 172 178 199 199 184 162 146 166
1952 171 180 193 181 183 218 230 242 209 191 172 194
1953 196 196 236 235 229 243 264 272 237 211 180 201
1954 204 188 235 227 234 264 302 293 259 229 203 229
1955 242 233 267 269 270 315 364 347 312 274 237 278
1956 284 277 317 313 318 374 413 405 355 306 271 306
1957 315 301 356 348 355 422 465 467 404 347 305 336
1958 340 318 362 348 363 435 491 505 404 359 310 337
1959 360 342 406 396 420 472 548 559 463 407 362 405
1960 417 391 419 461 472 535 622 606 508 461 390 432
```

3) Locus.ID – just id? no other logic?


Allele for short hair
Allele for white fur
Homologous pair of chromosomes
Locus for coat length gene
Locus for coat color gene
Allele for long hair
Allele for black fur

4) Where to find more examples “how to work with datasets”?
5) I don’t really understand the structure of the dataset. I guess it is a table, but how to get column names (what is in each row?)?
6) Should I learn more about Bioconductor? (https://www.bioconductor.org/help/course-materials/)

HAPNEST

HAPNEST

- **a novel approach for efficiently generating diverse individual-level genotypic and phenotypic data**.
- **a user-friendly tool** for generating **synthetic datasets for genotypes and phenotypes**, **evaluating synthetic data quality**, and **analysing the behavior of model parameters** with respect to the evaluation metrics.
- simulates genotypes by **resampling a set of existing reference genomes**, according to a **stochastic model** that approximates the underlying processes of coalescent, recombination and mutation
- similar to **HAPGEN2**
- enables **efficient simulation** of **diverse biobank-scale datasets**
- evaluating synthetic data **fidelity** and **generalisability**
- **approximate Bayesian computation (ABC)** techniques for analysing the **posterior distributions of model parameters** to aid model selection
- uses an **approximate model** inspired by the **sequential Markovian coalescent model**
-

Advantages:

5. Faster computational speed
6. Lower degree of relatedness with reference panels – (what does it mean???)
7. Generating datasets that preserve key statistical properties of real data

Key features:

- 6.8 million common variants and 9 phenotypes with varying degrees of heritability and polygenicity across 1 million individuals.
- focus on **reference-based approaches** (as PRSs we are mostly interested in common genetic variation)
- Synthetic haplotypes are constructed as a mosaic of segments of various lengths imperfectly copied from real haplotypes

7 methods to generate polygenic risk scoring across multiple ancestry groups and different genetic architectures:

1. MegaPRS
2. LDpred
3. Lassosum
4. PRScs
5. SBayesR
6. pT+clump
7. DBSLMM

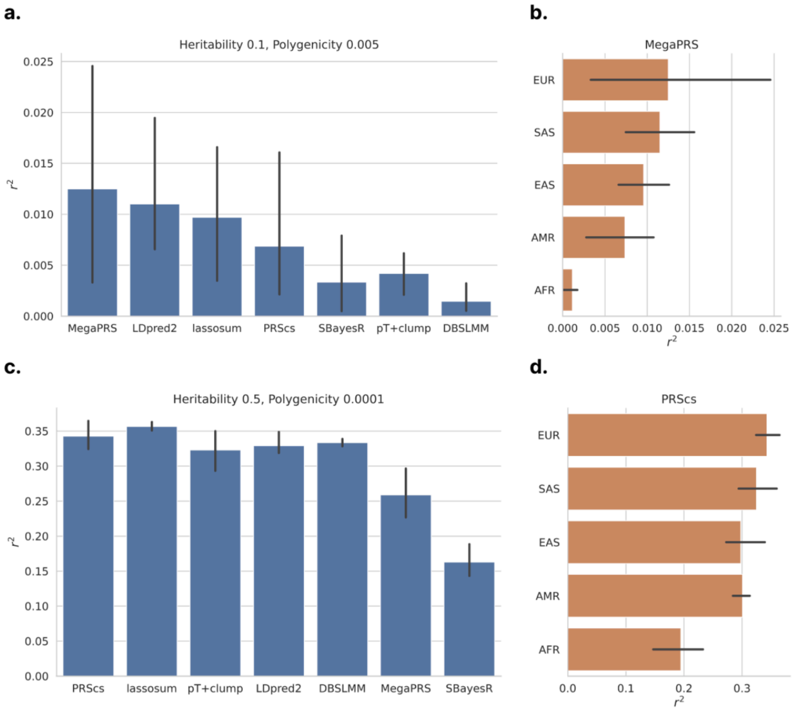


Figure 8: PRS results for two genetic architectures, averaged across 3 experiment trials with error bars showing the range of outcomes, for HapMap3 variants across 22 chromosomes. **a.** Pearson correlation between predicted and observed values, for various PRS methods and a European-ancestry phenotype with heritability 0.1 and polygenicity 0.005. **b.** Pearson correlation for various target ancestry groups for the best-performing PRS method (MegaPRS) for the heritability 0.1 and polygenicity 0.005 phenotype. **c.** Pearson correlation between predicted and observed values, for various PRS methods and a European-ancestry phenotype with heritability 0.5 and polygenicity 0.0001. **d.** Pearson correlation for various target ancestry groups for the best-performing PRS method (PRScs) for the heritability 0.5 and polygenicity 0.0001 phenotype.

2 main approaches have been used to simulate individual level genetic data:

1. Coalescence-based methods, such as Hudson's ms and msprime
   a. use demographic models to generate genomes
   b. including both rare and common variants.
2. **Reference-based approaches**
   a. use real genomic to generate synthetic data
   b. not suitable to generate realistic rare variants.

Reference-based approaches:

- **simGWAS** -- they do not meet modern demands for methods development based on individual level data. -> GWAS summary statistics
- **Hapmap3 SNPs** - are widely recommended for PRS computation
- **HAPGEN2** -- is a widely used tool for genotype and phenotype simulation, which preserves linkage disequilibrium (LD) patterns of real data through a resampling approach based on the Li and Stephens model. Lacks computational scalability and flexibility to simulate certain scenarios of interest for biobank-scale PRS and SNP-based methods development.
- **Sim1000G** is an integrated R package, but is limited to genotype simulation.
- **G2P** encompasses both genotype and phenotype simulation, and is highly customisable, but this setup can be challenging for non-expert users.

Differences from **HAPGEN2:**

- The simulation varying, rather than constant, coalescence time T
- The presence of a genetic variant at position i is only copied if T ≤ mi, where mi is the variant's age of mutation

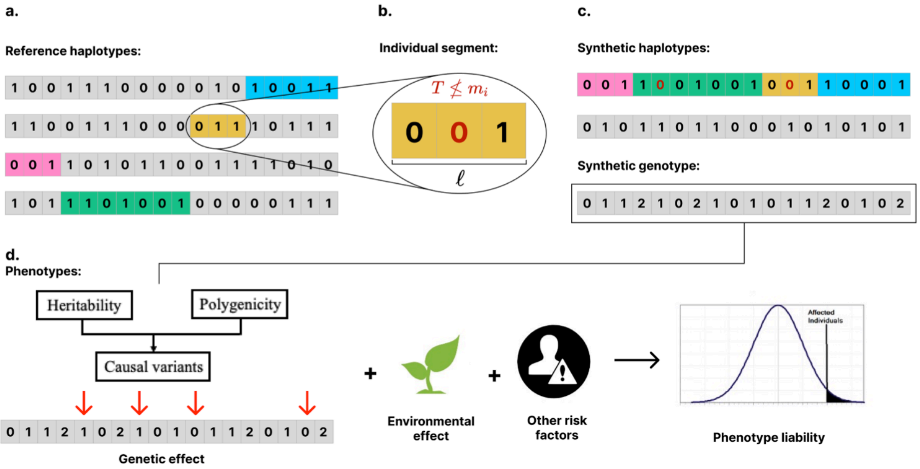


Figure 1: **a.** A reference set of real haplotypes, from which segments (colored) are imperfectly copied to construct a synthetic haplotype. **b.** Detailed view of an individual segment. The segment length, ℓ, and coalescence time, T, are sampled from a stochastic model. The presence of a genetic variant at position i is only copied if $T \leq m_i$, where m_i is the variant's age of mutation. Variants that are not copied are shown in red. **c.** Synthetic genotypes, g, are constructed as pairs of synthetic haplotypes, h_i, $i \in \{1, 2\}$. **d.** Once the genotype is generated, liability of phenotype will subsequently be assigned to each sample as a summation of genetic effect, covariate effect (if any) and environmental noise.

Fidelity measurement as the similarity between the real (reference) and synthetic datasets for 4 properties:

- minor allele frequency (MAF) distribution,
- population structure in terms of alignment of the principal components (PCs),
- LD decay
- nearest neighbour adversarial accuracy

evaluating synthetic data quality – how in HAPNEST?
evaluating methods for polygenic risk score – how in HAPNEST?

Dictionary:

SNP - single nucleotide polymorphisms
PRS - polygenic risk scoring (study: https://www.nature.com/articles/s41596-020-0353-1)
GWAS - genome-wide association studies
LD - linkage disequilibrium
ABC - approximate Bayesian computation
MAF - minor allele frequency
PC - principal components (PCs)

Takes:

- *…the development of methods that can improve the generalisability of PRSs is needed …*
- *Without an integrated approach for parameter selection and evaluation of synthetic data quality, **it is difficult for end-users to understand the statistical guarantees and reliability** of the generated datasets.*
- *…there **does not exist a software tool** implementing an end-to-end pipeline **for synthetic data generation, evaluation and optimisation**. (before* HAPNEST *)*

Questions:

1) polygenic risk score – what is it?
2) lower degree of relatedness with reference panels – what does it mean?
3) … nine phenotypes… Only nine?
4) evaluating synthetic data quality – how in HAPNEST?
5) evaluating methods for polygenic risk score – how in HAPNEST?
6) Julia code (https://github.com/intervene-EU-H2020/synthetic_data). Study Julia lang?
7) How do we get real haplotypes? Is it legal?
   …we consider a reference dataset of 4,062 phased genotypes derived from the publicly available 1,000 Genomes Project and Human Genome Diversity Pro ject datasets for 6 major discrete ancestry groups *… ???*
8) Why these distributions (https://en.wikipedia.org/wiki/Exponential_distribution, https://en.wikipedia.org/wiki/Gamma_distribution)? Because we need just 1 mutation in l and T referees to *k* events that should happen for mutation (we wait until age/time T)?

$$\ell \sim Exp(2T\rho_s),\ T \sim Gamma(2, N_s/N_{e,s}),$$

9) I know principal components from PCA. Are they different here?

Выводы:

1) Как это работает? – пока что в процессе понимания
2) Не помешают ли внутренние предположения, на которых основана эта модель, тому, чтобы любые полученные результаты по анализу таких данных не были полностью артефактом способа генерации данных? – Думаю, что можно получить хорошие результаты, но стоит обратить вниание на следующее:
   - focuses on **reference-based approaches => not suitable to generate realistic rare variants.**
   - *…However, we would like to note that this approach does not accurately reflect the process of multi-population diverging and intermixing, therefore it should be used and interpreted carefully…*
   - *…we note that the criteria used in our analysis are not sufficient for differential privacy guarantees, and **we advise to use HAPNEST, or any of the reference-based generation methods, only on publicly-available genomics datasets**…*

TODO

- Finish HAPNEST paper studying
- https://www.ebi.ac.uk/biostudies/studies/S-BSST936. What to download and how to work?
- Learn more about Bioconductor
- Study: https://www.nature.com/articles/s41596-020-0353-1
- Study Julia lang (?)