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

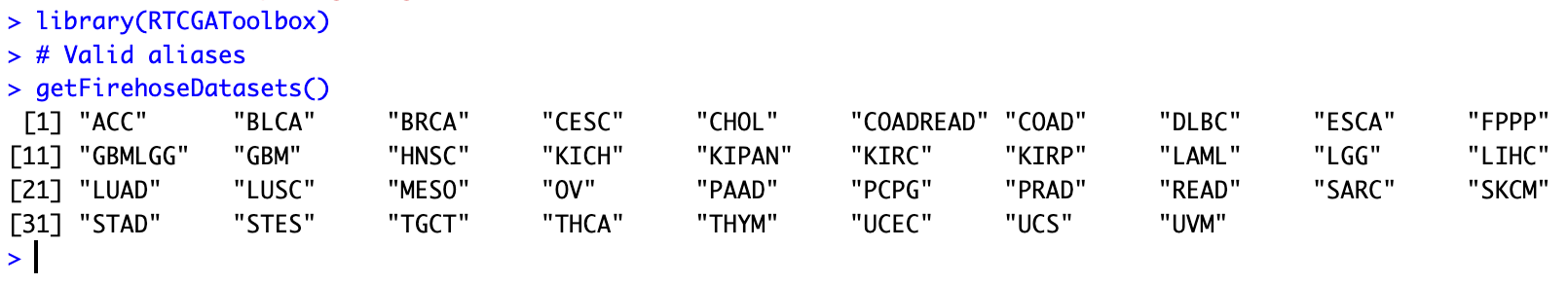
1. **Check valid dataset aliases, stddata run dates and analyze run dates:**

getFirehoseDatasets() – valid dataset aliases

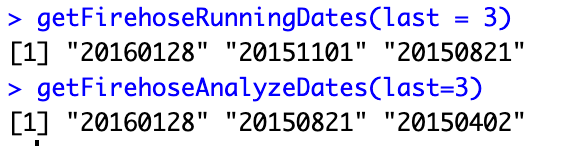
getFirehoseRunningDates() – stddata run dates – date of data generation

getFirehoseAnalyzeDates() – analyze run dates -- date of data analysis

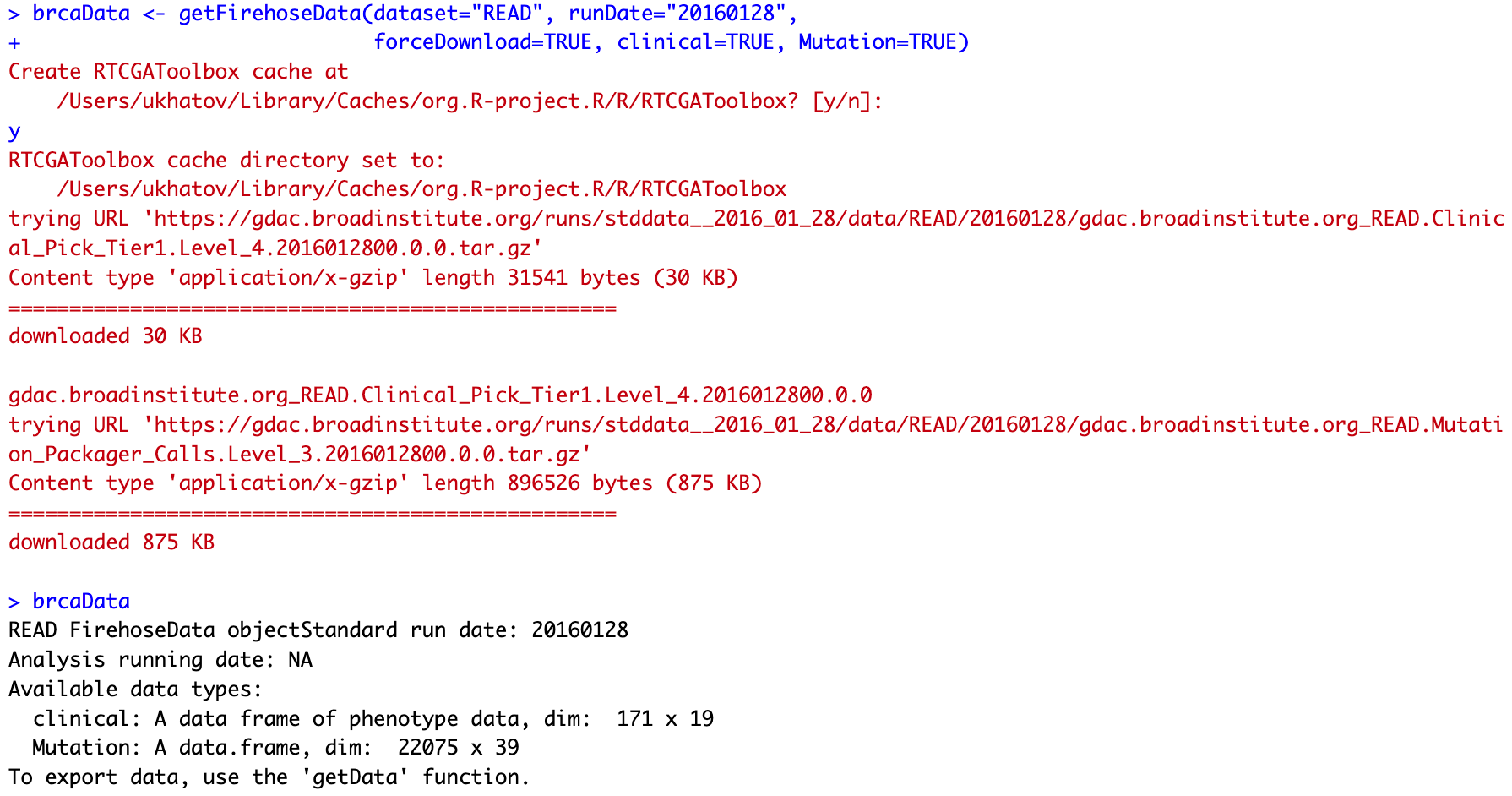
|  |
| --- |
| **library**(RTCGAToolbox)  *# Valid aliases*  getFirehoseDatasets() |



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



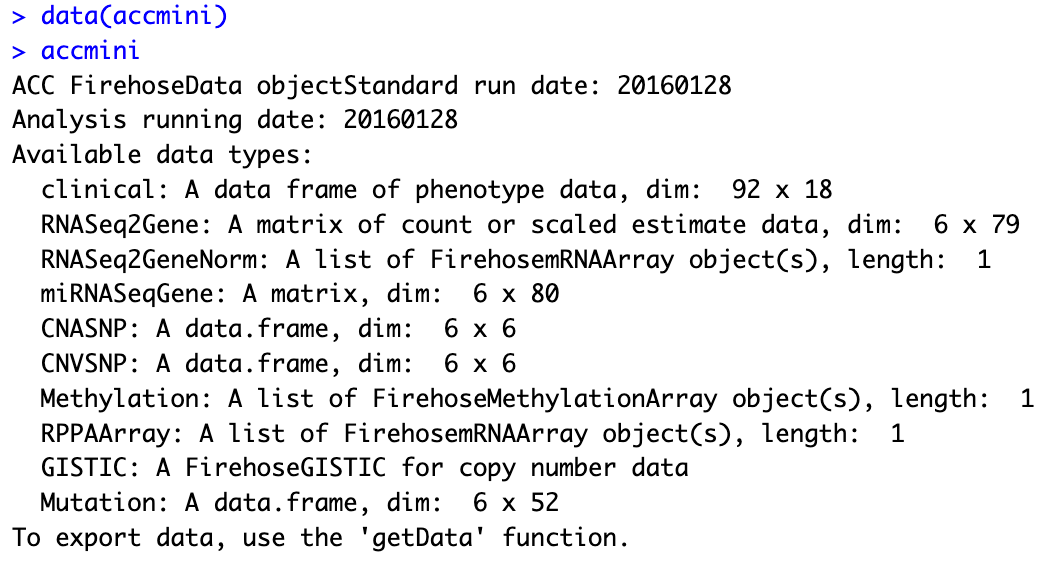
|  |
| --- |
| *# READ mutation data and clinical data*  brcaData <- getFirehoseData(dataset="READ", runDate="20160128",  forceDownload=TRUE, clinical=TRUE, Mutation=TRUE)  brcaData |



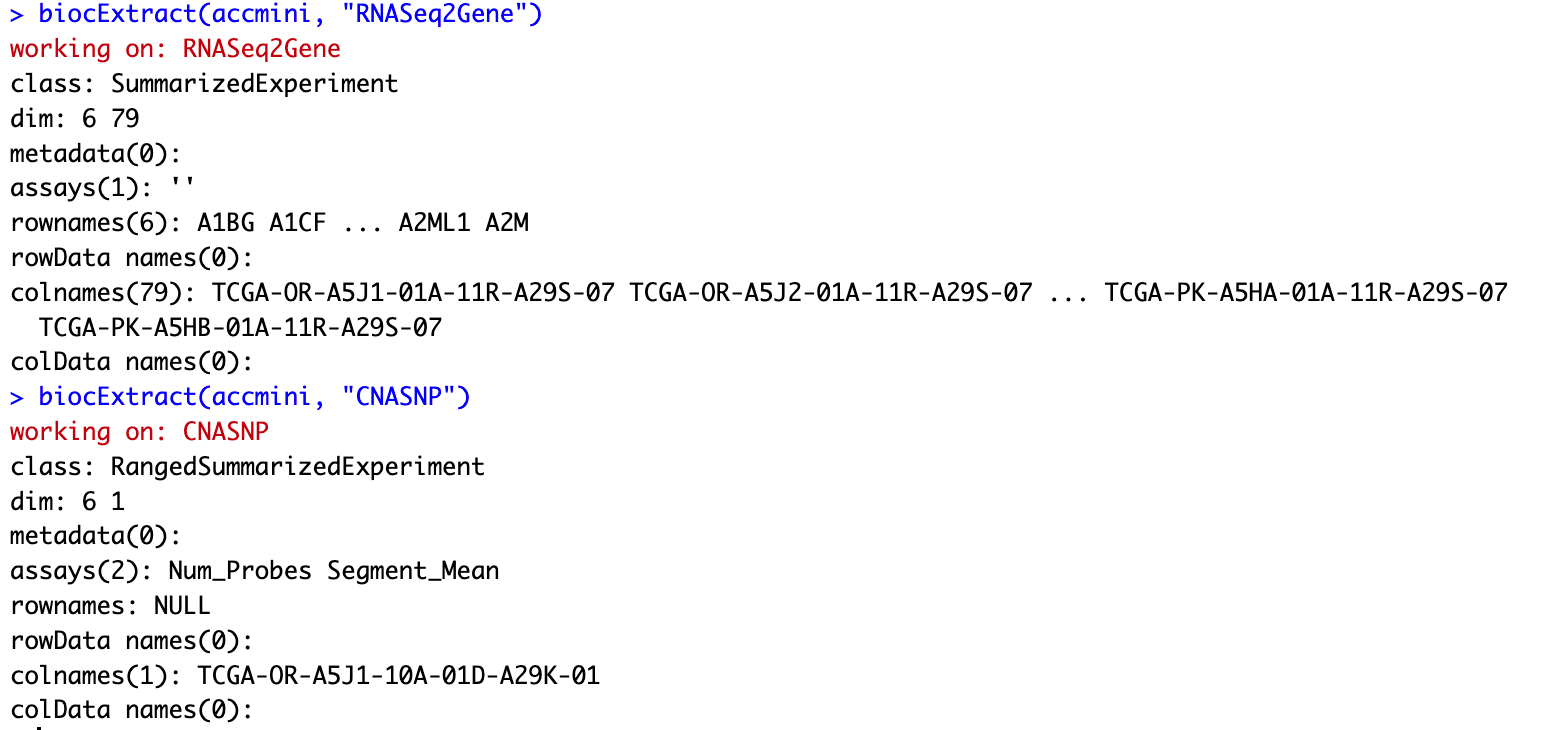
1. **Example Dataset Exploration**

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

|  |
| --- |
| data(accmini)  accmini |



|  |
| --- |
| biocExtract(accmini, "RNASeq2Gene")  biocExtract(accmini, "CNASNP") |

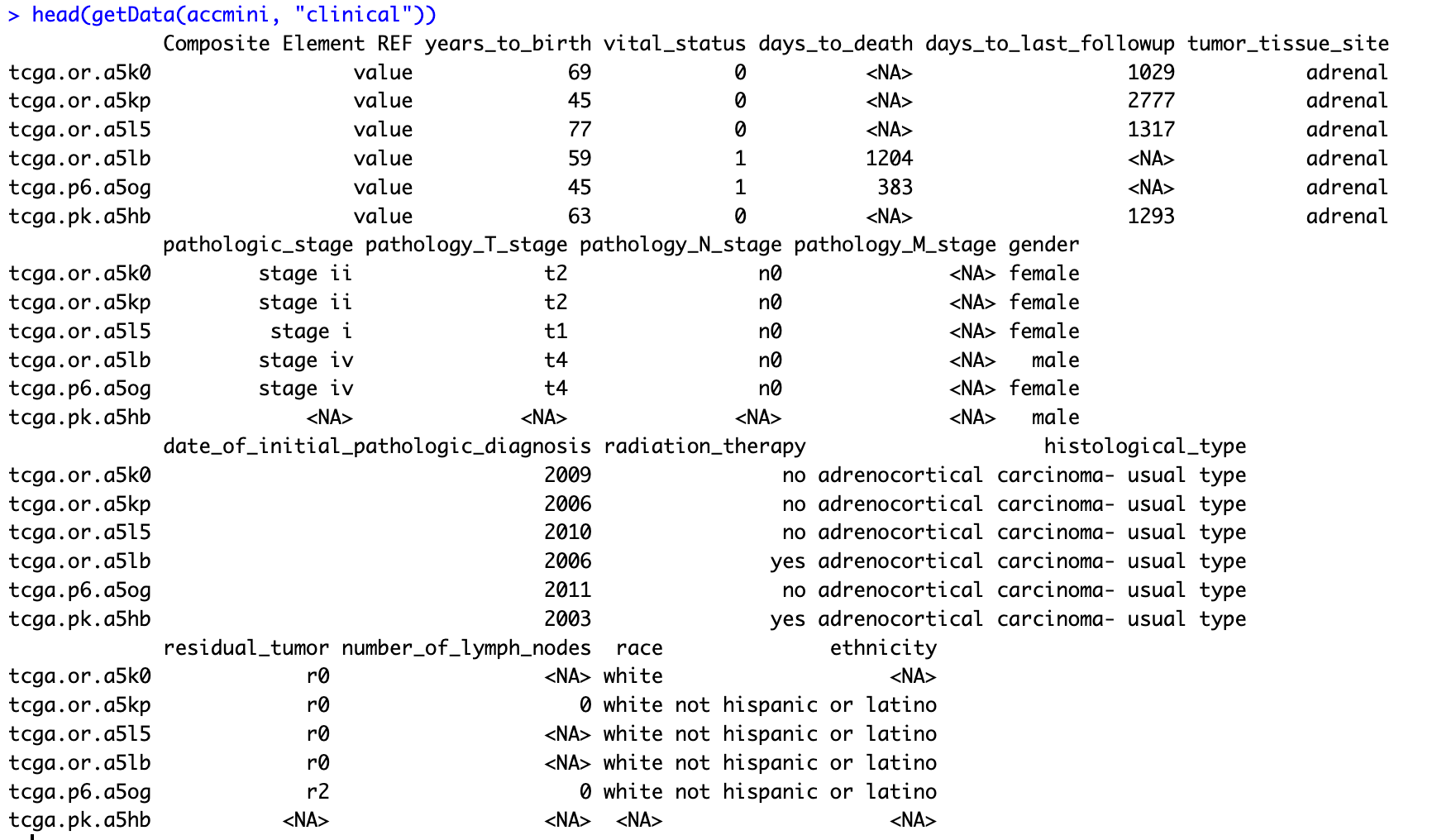


Following logic keys are provided for different data types. By default client only download clinical data.

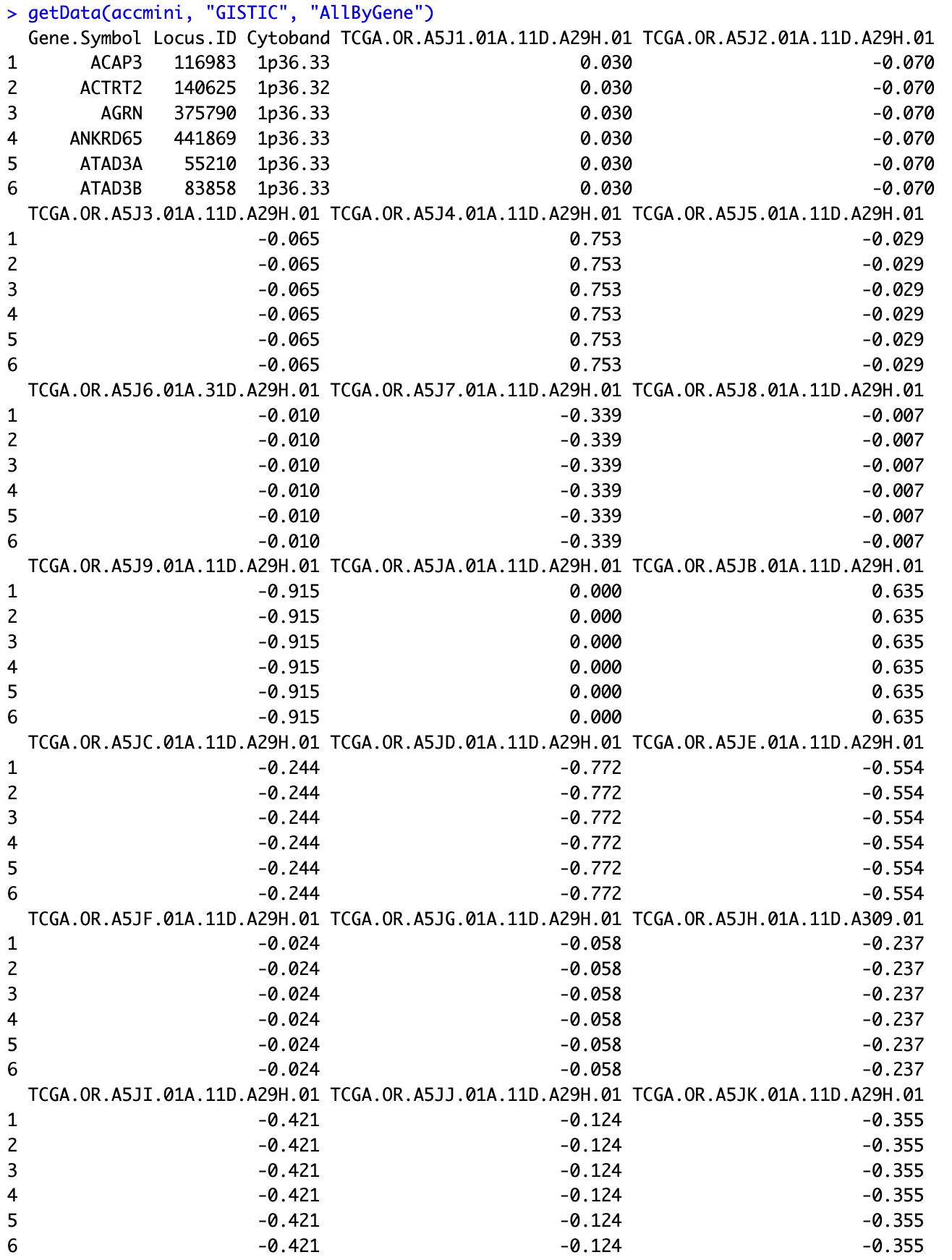
* RNAseqGene
* clinical
* RNASeqGene
* RNASeq2Gene
* RNASeq2GeneNorm
* miRNASeqGene
* CNASNP
* CNVSNP
* CNASeq – copy number altaration seq (пол хромосомы удвоилось или потерялось, …)
* CNACGH
* Methylation
* Mutation
* mRNAArray
* miRNAArray
* RPPAArray

1. **Raw Data**

|  |
| --- |
| head(getData(accmini, "clinical"))  getData(accmini, "RNASeq2GeneNorm")  getData(accmini, "GISTIC", "AllByGene") |

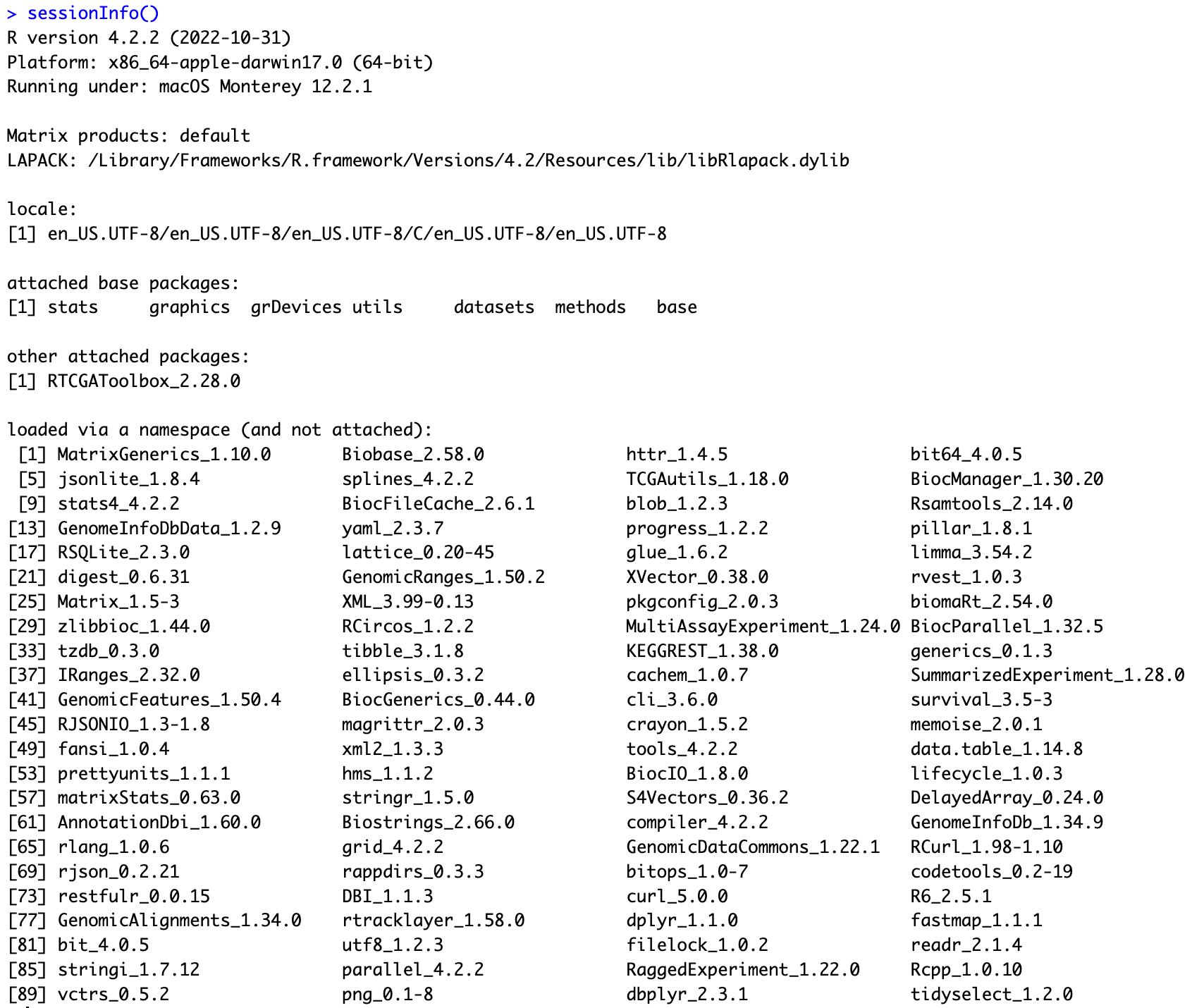






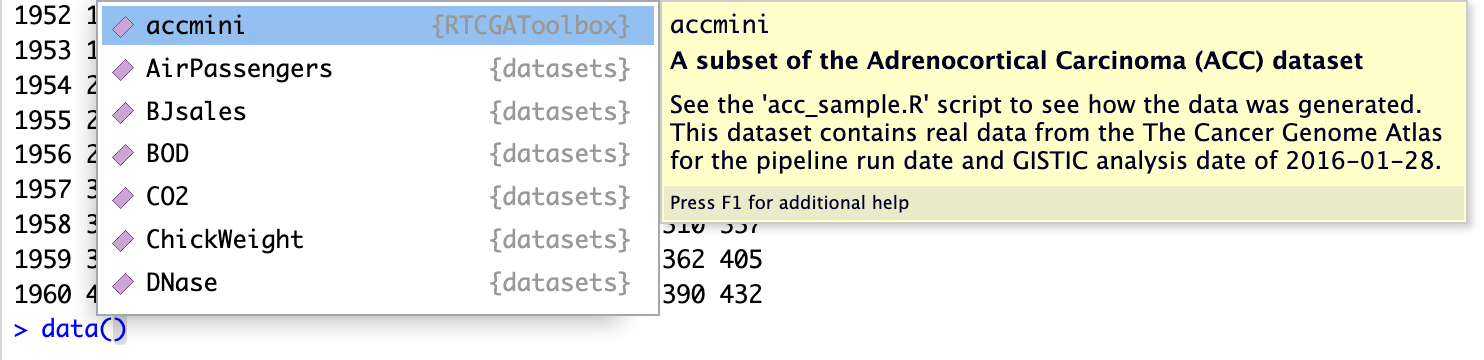
1. **Session information**

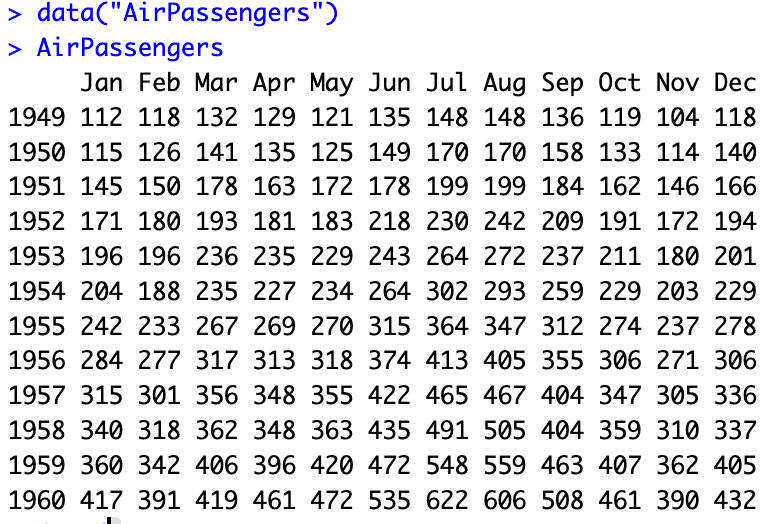
|  |
| --- |
| sessionInfo() |



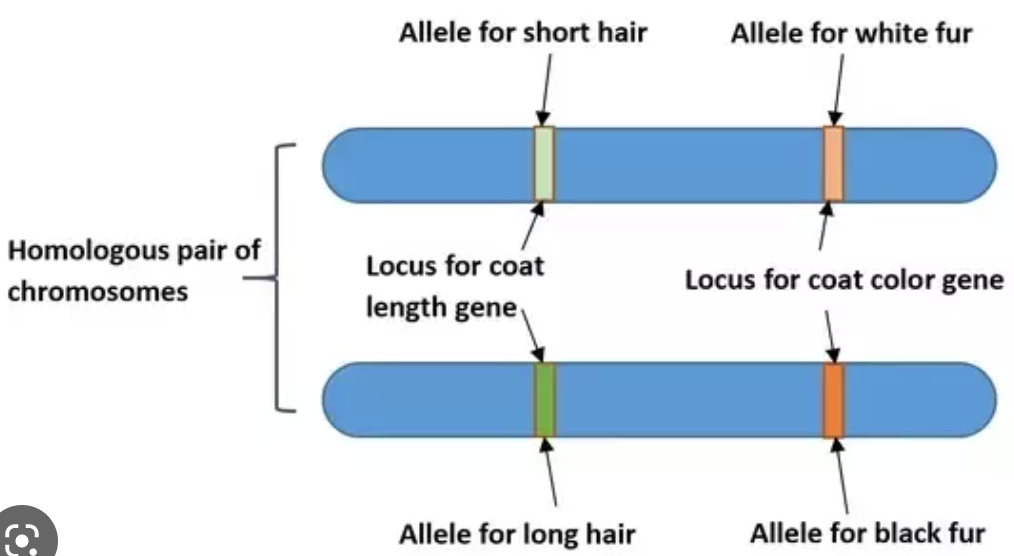
Questions:

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…





1. Locus.ID – just id? no other logic? Gene ID в ncbi



1. Where to find more examples “how to work with datasets”?
2. 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?)?
3. Should I learn more about Bioconductor? (<https://www.bioconductor.org/help/course-materials/>) – **yes, but 1000 пакетов => слишком много всего, не тратить много времени -> гуглить побольше**

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:

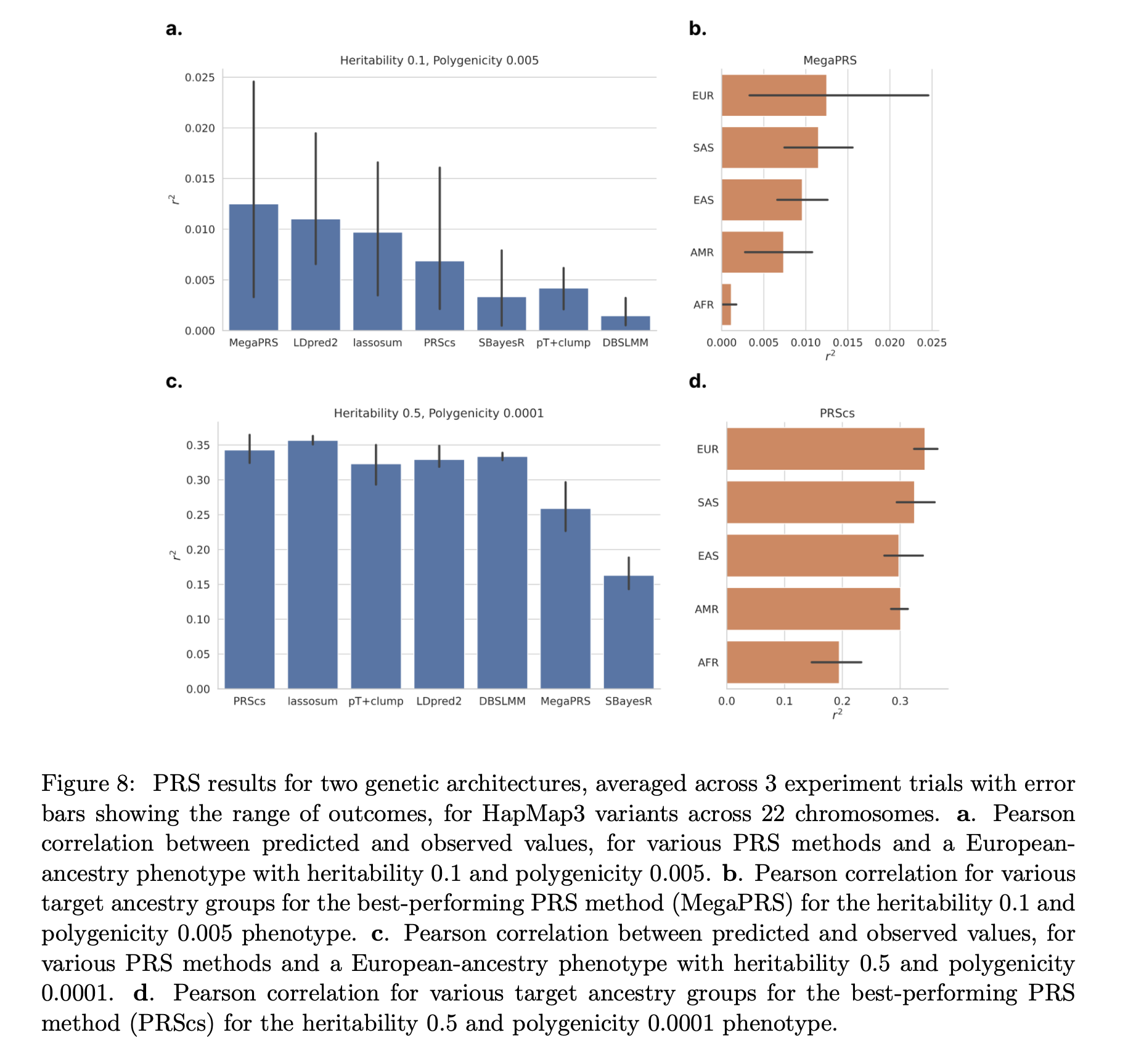
1. Faster computational speed
2. Lower degree of relatedness with reference panels – (what does it mean???)
3. Generating datasets that preserve key statistical properties of real data

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



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

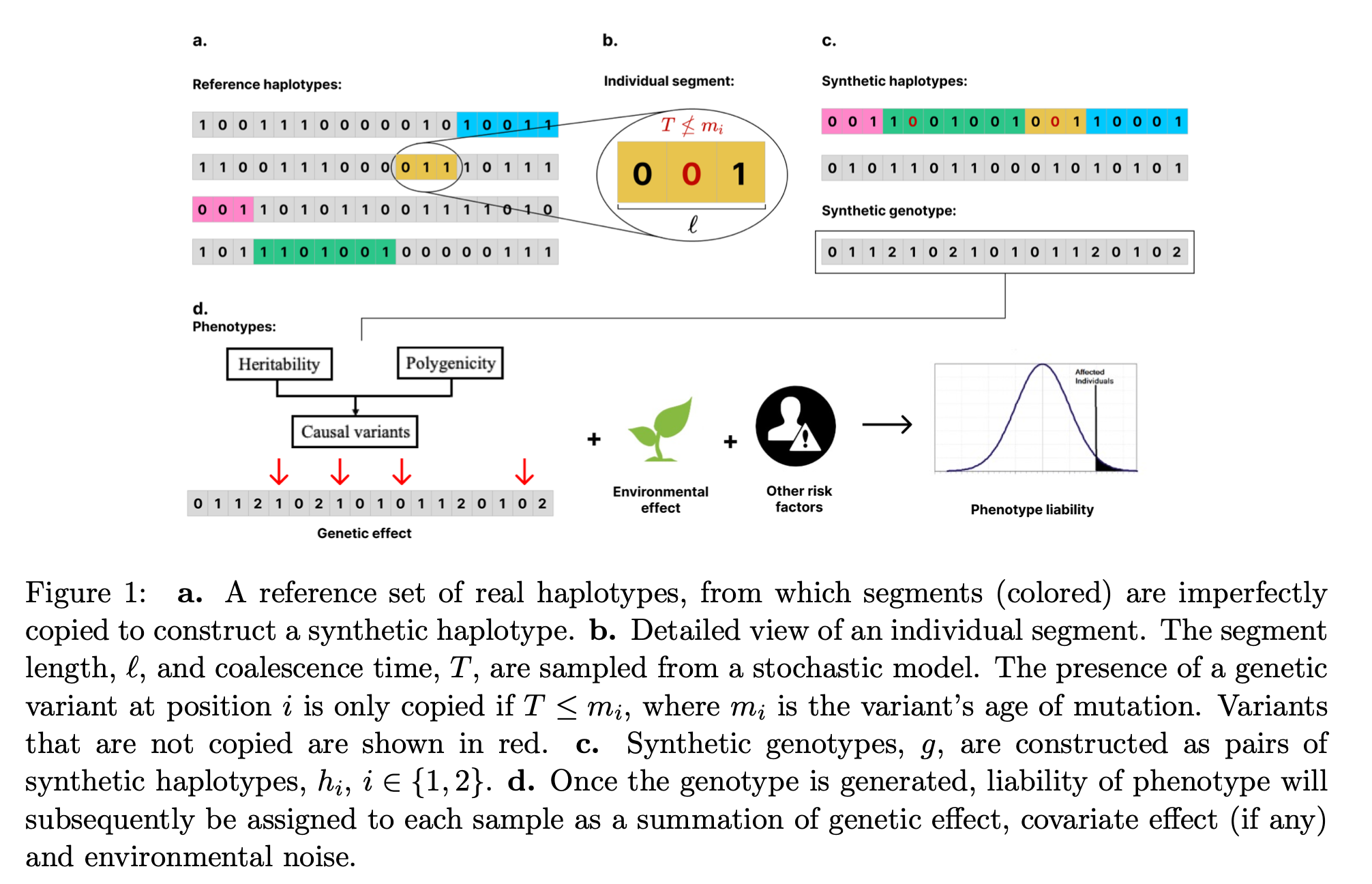
1. Coalescence-based methods, such as Hudson’s ms and msprime
   1. use demographic models to generate genomes
   2. including both rare and common variants.
2. **Reference-based approaches** 
   1. **use real genomic to generate synthetic data**
   2. **not suitable to generate realistic rare variants.**

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



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:

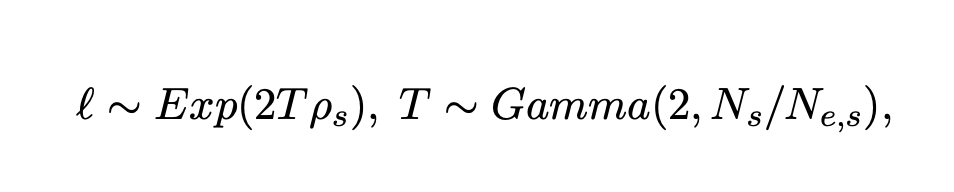
* …*the development of methods that can improve the generalisability of PRSs is needed …*
* *Without an integrated approach for parameter selection and evaluation of synthetic data quality,* ***it is difficult for end-users to understand the statistical guarantees and reliability*** *of the generated 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 *… ???*

1. Why these distributions (<https://en.wikipedia.org/wiki/Exponential_distribution>, https://en.wikipedia.org/wiki/Gamma\_distribution)? Because we need just 1 mutation in and referees to *k* events that should happen for mutation (we wait until age/time *)?*



1. 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**

* "LGG" dataset (часто мутация в IDH1 – это хорошо или плохо). Разный уровень экспресии гена LGG – есть разница в выживаемости? **Survival Analysis** (**COXPH** (proportion hazzard ) = regression model (добавить ковариаты: возраст, тип опухоли, информация о типе именной системы, курят, лечение? ))
* 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 -> **no -> execute out of the box -> получить результат**

Иммунитет:  
- Главный комплекс … (MHC = HLA (в случае человека))   
  
Тип = каким образом нарезаются кусочки белков   
  
Иммунитет хороший = резать и показывать кусочек, который не мутирует!  
  
Нет меланоцита => нет меланом  
  
Знаю HLA = знаю кусочки, которые нужно показать иммунитету (прививка от рака)  
  
HLA закодирован с помощью ДНК   
  
кусок белка – на сколько иммуногенен?

HLA локус (на 6 хромосоме) – очень повторяющаяся последовательность. Сложно понять, из какого региона пришел кусок ДНК => сложно понять какой HLA по генам

**Algorithm optitype** – local alignment



Class III – also exists (search more)

<https://pubmed.ncbi.nlm.nih.gov/11977834/>

сгенерировать PRS (7 штук) с помощью dataset HAPNEST -> посмотреть на сколько разные?

Concordance (согласовать)

**Survival Analysis** (**COXPH** (proportion hazzard ) = regression model

**Silent VS [others]**

TCGA barcode - study  
  
TCGA-DU-6394-01A-11D-1705-08  
DU-6394 – unique identity number for the person  
01A-аликвота   
10< - mutation  
10+ -- норм  
Tumor\_Sample\_Barcode