TCGA Pan Cancer Data Analysis R Package

ETC 5543

Abhishek Sinha

31322743

Motivation

- The Cancer Genome Atlas (**TCGA**) is a large and complex project that collects tumor samples from different institutions and at different times.

- TCGA datasets are comprehensive and in-depth datasets organized from the analysis of over 11000 tumor samples from 33 of the most prevalent Cancer types.

- This makes TCGA Datasets one of the widely used datasets in Cancer research and is an essential resource for the development of new treatments.

- But given the complexity involved in gathering data it becomes prone to unwanted variation such as *batch effects* and *time effects*.

- This is a main challenge in the analysis of gene expression data as presence of unwanted variations leads to false positive or misleading biological conclusions resulting in retractions.

Solution

- One basic solution to the problem is normalization.
- Along with the raw count data TCGA provides two normalized datasets, FPKM and FPKM.UQ.
- But these normalization methods often fail to remove the unwanted variations.
- A novel approach to this problem is being proposed [1], using Pseudo Replicates of Pseudo Sample (PRPS), to deploy RUV – III normalization.
- To facilitate the application of methods used in identifying unwanted variations and applying RUV-III normalization method [1], a tool was needed.
- My role was to develop a R Package that would help to communicate the findings of the Paper [1] and help Bioinformaticians and Biologists with a tool that can be used for handling variations.

[1]. https://www.biorxiv.org/content/10.1101/2021.11.01.466731v1

TGCA Data:

- The data used for the analysis is TCGA RNA-Seq data for Breast Cancer.
- The data is loaded using ***SummarizedExperiment*** class which is package in *Bioconductor*. This package is a matrix like data container where rows have information about Genes, columns represent information about Samples and objects contains one or more assays.
- In TGCA Breast Cancer Dataset there are:

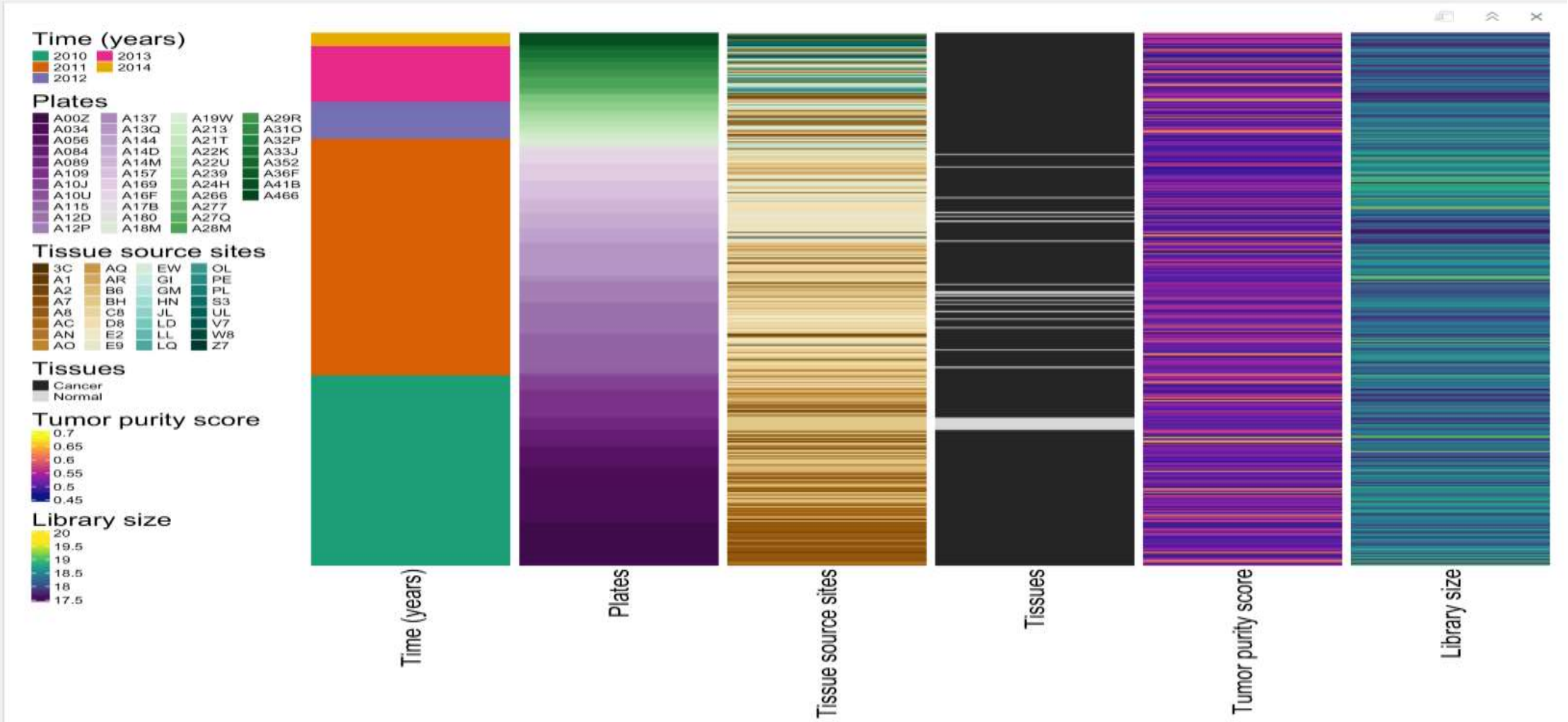
1. Gene Data with 56,493 observations and 40 features.
2. Sample Data with 1,222 observations and 4,115 features.
3. Assays for raw count, FPKM and FPKM.UQ with 56,493 observations and 1,222 sample features.

R Package – *tcgapkg* functionality

- The Package contains a subset of original data to run the package functions.
- The R Package also contains multiple filtering functions that help to filter data based on user requirements.
- This includes functions to filter data by gene type, remove lowly expressed gene, Tumor Purity and Library Size.
- The Package also contains set of Data Analysis functions that help identify variations and run RUV-III.
- I added Vignette as a guide to the user on different functionalities of the package and how to use them

Study Design

```
86 # Data Analysis
87
88 # Study Design Plot
89
90 ```{r fig.align='center', fig.width=7, fig.height=5}
91 study.design(data = df4)
92 ```
```



PCA

- It is the primary function to identify unwanted variation.
- The PCA function generates PCs using BiocSingular::SVD algorithm.
- Benefit of using this algorithm is the significantly reduced processing time.
- User needs to supply the data and number of PCs required.

```
# PCA

## Generate PCA

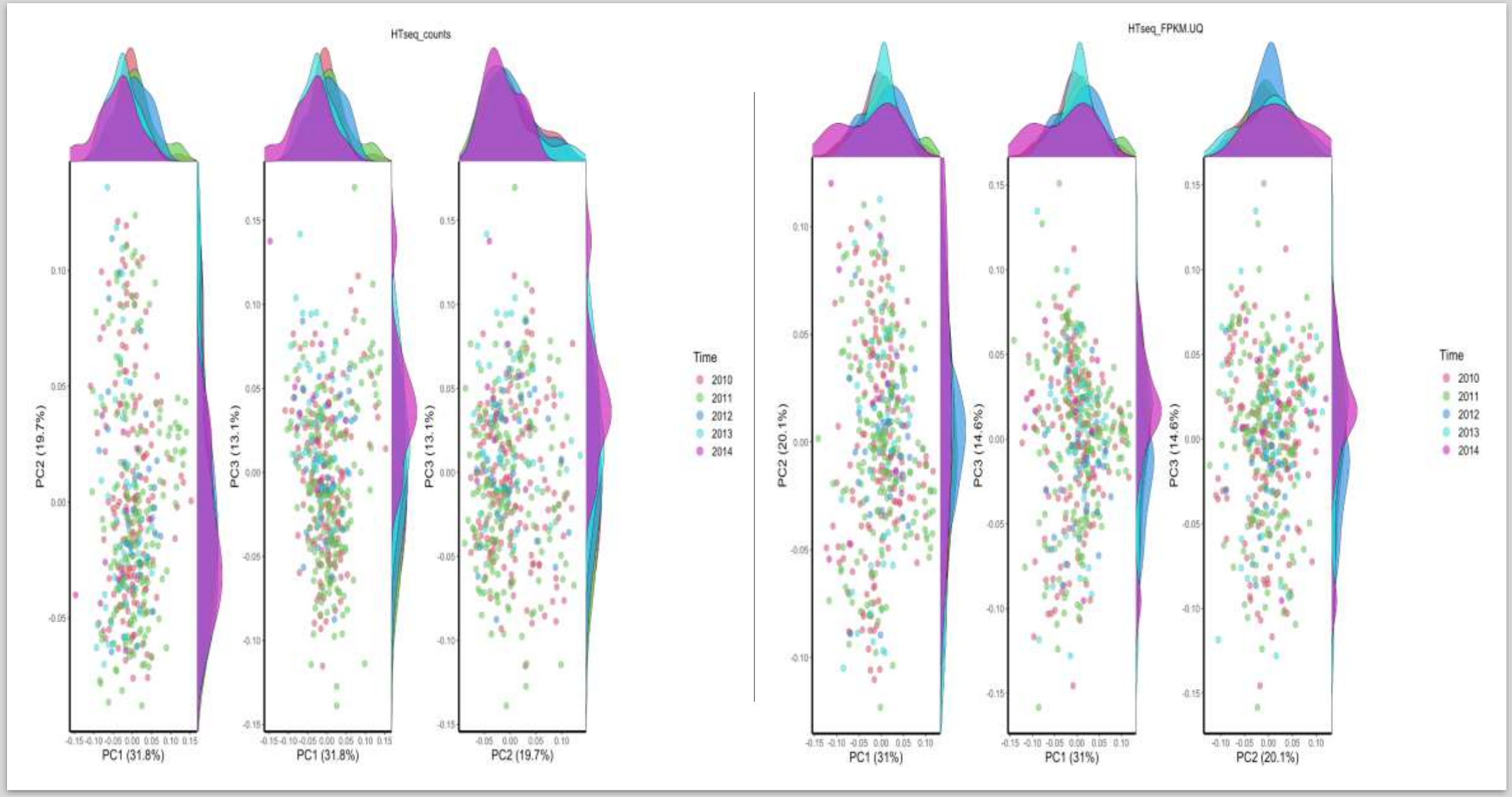
```{r}
# Is data input for PCA logical
is.logical(df4)
```

[1] FALSE

```{r}
df5 <- get.pca(data = df4, nPcs = 7, is.log = FALSE)
```
```

```
## Plot PCA

```{r fig.align='center', fig.width=7, fig.height=5, message=FALSE, warning=FALSE}
library(ggplot2)
library(cowplot)
pca.plot(pca.data = df5, data = df4, group = "Time", plot_type = "DensityPlot", npcs = 3)
```
```

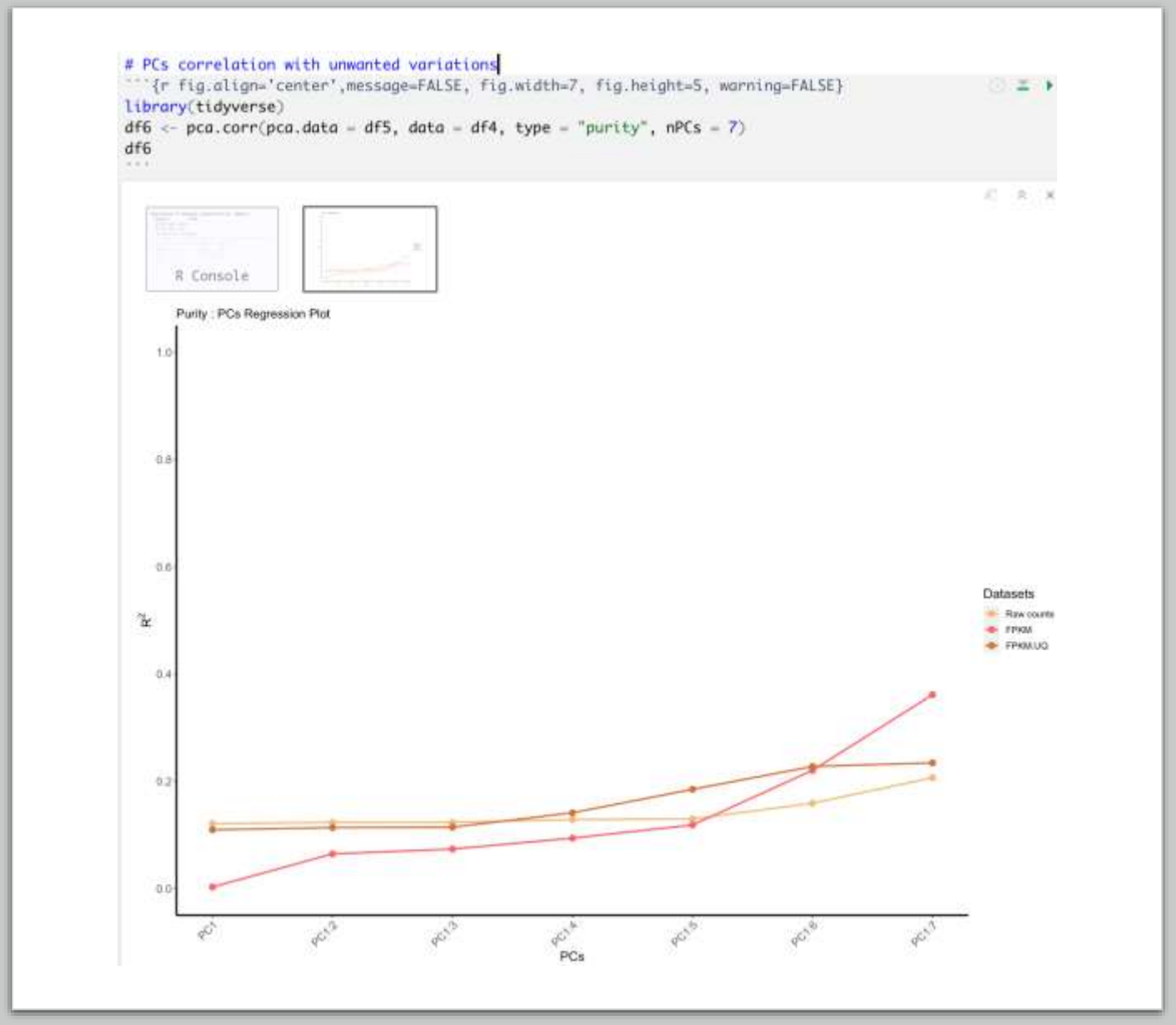

HTseq_counts
HTseq_FPKM.UQ
PC1 (31.8%)
PC2 (19.7%)
PC3 (13.1%)
PC1 (31%)
PC2 (20.1%)
PC3 (14.6%)
Time
2010
2011
2012
2013
2014

Regression - Vector Correlation

- Linear Regression is used to quantify the relationship between the first few PCs and continuous sources of unwanted variation such as (log) library size , Purity score.
- The R^2 calculated from the linear regression analyses indicates how strongly the PCs capture unwanted variation in the data.
- Similarly, to linear regression, we used vector correlation analysis to assess the effect on the data of discrete sources of unwanted variation such as years or year intervals.

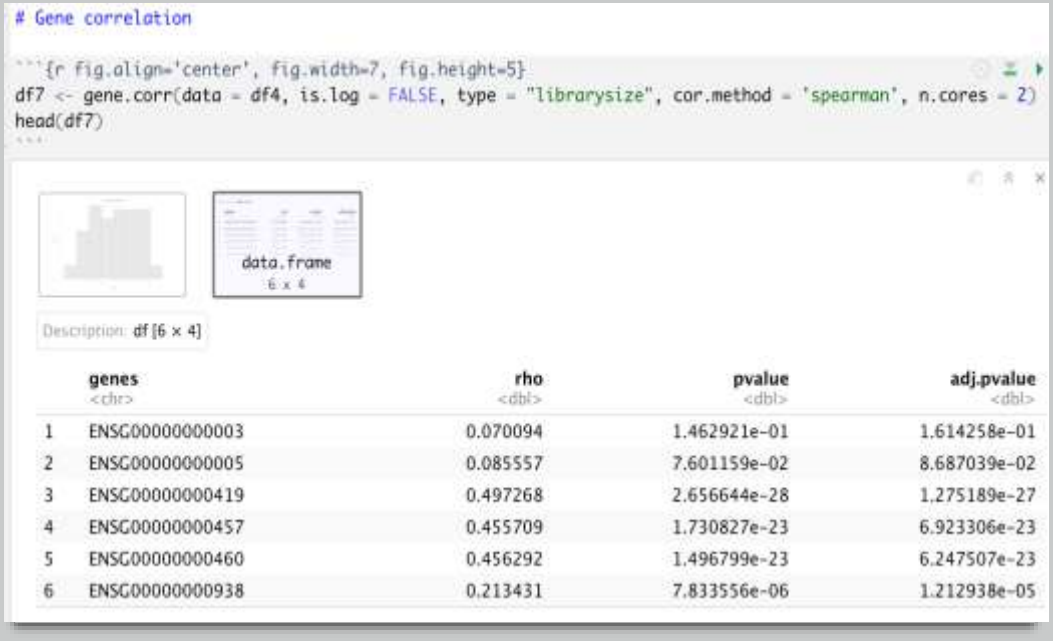
R^2 Plot

- The output of the plot includes linear plot for each assay which plots the R^2 values across PCs.
- In this plot if there is unwanted variation in the data, we will see PCs with high R^2 values.
- Looking at this plot we can see that FPKM.UQ has not handle the variations properly.
- FPKM normalization does better job in handling variation.



Statistical Tests

- To add to the regression analysis, we can perform statistical tests to further explain the relation between variations and genes.
- Spearman Correlation test can be used to understand the degree of correlation between Library size and individual genes in raw data.
- We use ANOVA F statistics to summarize the effects of a qualitative source of unwanted variation (e.g., batches) on the expression levels of individual genes.
- Genes having large F -statistics are deemed to be affected by the unwanted variation.



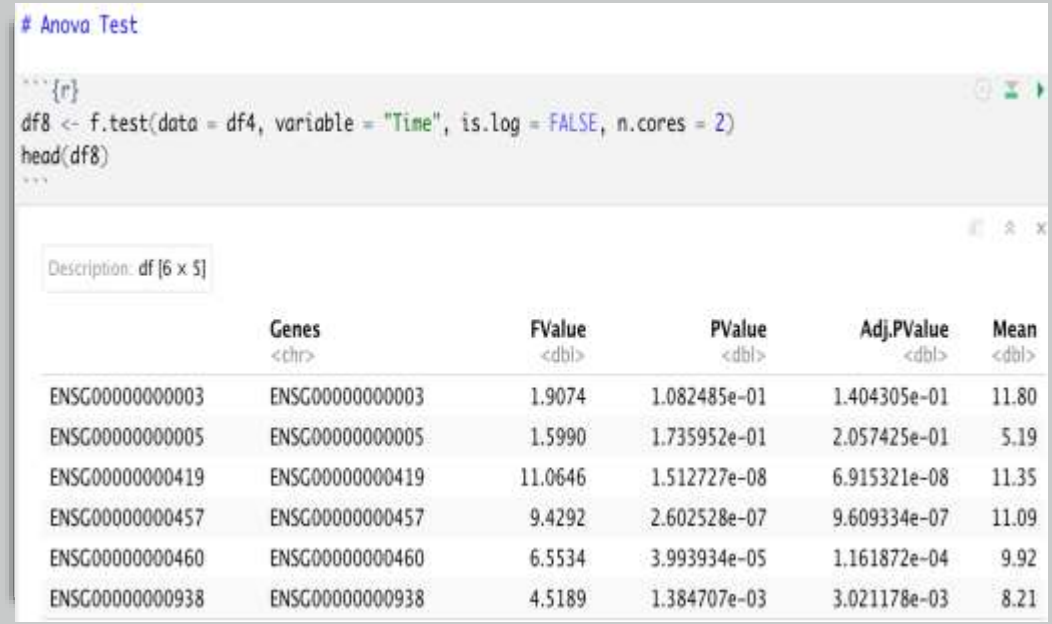
```
# Gene correlation

```{r fig.align='center', fig.width=7, fig.height=5}
df7 <- gene.corr(data = df4, is.log = FALSE, type = "librarysize", cor.method = 'spearman', n.cores = 2)
head(df7)
```
```

data.frame
6 x 4

Description: df [6 × 4]

	genes <chr>	rho <dbl>	pvalue <dbl>	adj.pvalue <dbl>
1	ENSG00000000003	0.070094	1.462921e-01	1.614258e-01
2	ENSG00000000005	0.085557	7.601159e-02	8.687039e-02
3	ENSG00000000419	0.497268	2.656644e-28	1.275189e-27
4	ENSG00000000457	0.455709	1.730827e-23	6.923306e-23
5	ENSG00000000460	0.456292	1.496799e-23	6.247507e-23
6	ENSG00000000938	0.213431	7.833556e-06	1.212938e-05



```
# Anova Test

```{r}
df8 <- f.test(data = df4, variable = "Time", is.log = FALSE, n.cores = 2)
head(df8)
```
```

Description: df [6 × 5]

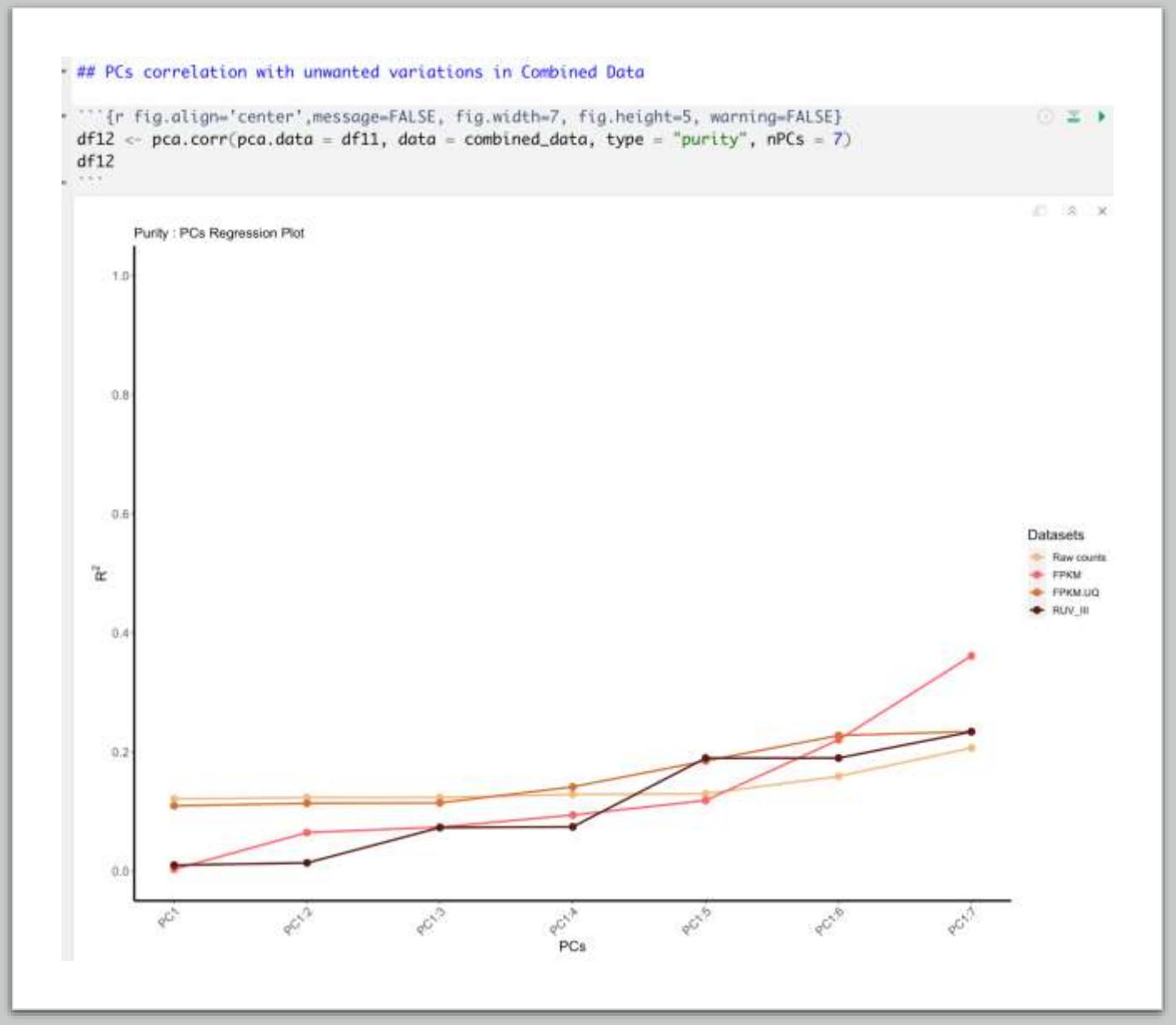
	Genes <chr>	FValue <dbl>	PValue <dbl>	Adj.PValue <dbl>	Mean <dbl>
ENSG00000000003	ENSG00000000003	1.9074	1.082485e-01	1.404305e-01	11.80
ENSG00000000005	ENSG00000000005	1.5990	1.735952e-01	2.057425e-01	5.19
ENSG00000000419	ENSG00000000419	11.0646	1.512727e-08	6.915321e-08	11.35
ENSG00000000457	ENSG00000000457	9.4292	2.602528e-07	9.609334e-07	11.09
ENSG00000000460	ENSG00000000460	6.5534	3.993934e-05	1.161872e-04	9.92
ENSG00000000938	ENSG00000000938	4.5189	1.384707e-03	3.021178e-03	8.21

PRPS

- Pseudo Replicates of Pseudo Samples is a method proposed to handle the problem of missing technical replicates in TCGA Datasets.
- Since RUV-III method is based on the concept of technical replicates it becomes a challenge if the data does not have it.
- The gene expression measurements of biologically homogeneous sets of samples are averaged within batches, and the results called pseudo-samples.
- Pseudo-samples with the same biology and different batches are then defined to be pseudo-replicates.
- The variation between pseudo-samples of a set pseudo-replicates is mainly unwanted variation.

RUV - III

- RUV-III is a previously developed method by my Supervisor [2], that uses negative controlled genes and technical replicates to estimate variation.
- RUV-III normalizes the differences between two or more pseudo-samples with the same biology.
- Negative controls for RUV-III are genes that have reasonable expression levels and whose variation is largely unwanted, i.e., not of biological interest.



[2]: https://minerva-access.unimelb.edu.au/handle/11343/226888

Future Work:

- The package website that can be easily accessible and provides information about package, individual functions and articles on package use.
- Adding unit tests to add additional checks to package performance.