PCAGO: An interactive web service to analyze RNA-Seq data with principal component analysis

Ruman Gerst, Manja Marz and Martin Hölzer

33rd Winterseminar Bled

February 14, 2018

Friedrich Schiller University Jena
RNA Bioinformatics and High-Throughput Analysis

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Martin Hölzer

presents

FUN with PCA!

Principal Component Analysis

Method that takes a dataset with a lot of dimensions (i.e. lots of RNA-Seq samples) and flattens it to 2 or 3 dimensions so we can look at it.

It tries to find a meaningful way to flatten the data by focusing on the things that are most different (most variant) between samples.

Principal Component Analysis

Method that takes a dataset with a lot of dimensions (i.e. lots of RNA-Seq samples) and flattens it to 2 or 3 dimensions so we can look at it.

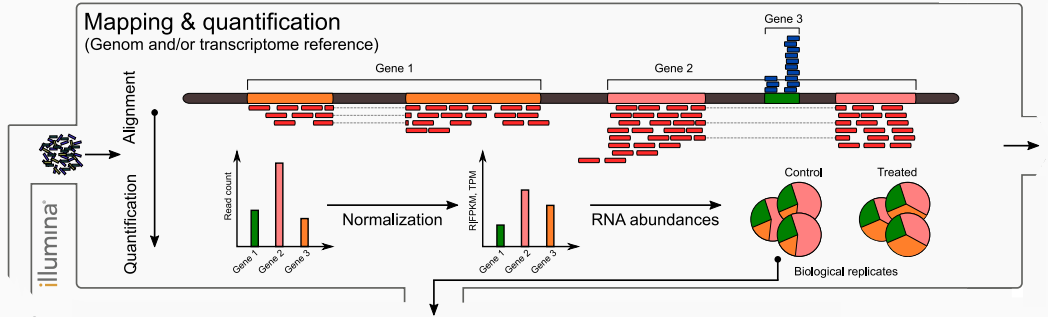
It tries to find a meaningful way to flatten the data by focusing on the things that are most different (most variant) between samples.

Principal Component Analysis

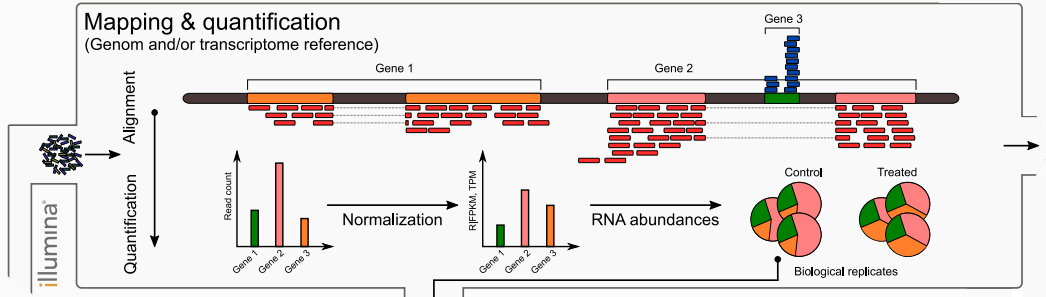
Method that takes a dataset with a lot of dimensions (i.e. lots of RNA-Seq samples) and flattens it to 2 or 3 dimensions so we can look at it.

It tries to find a meaningful way to flatten the data by focusing on the things that are most different (most variant) between samples.

Reference-based RNA-Seq analysis



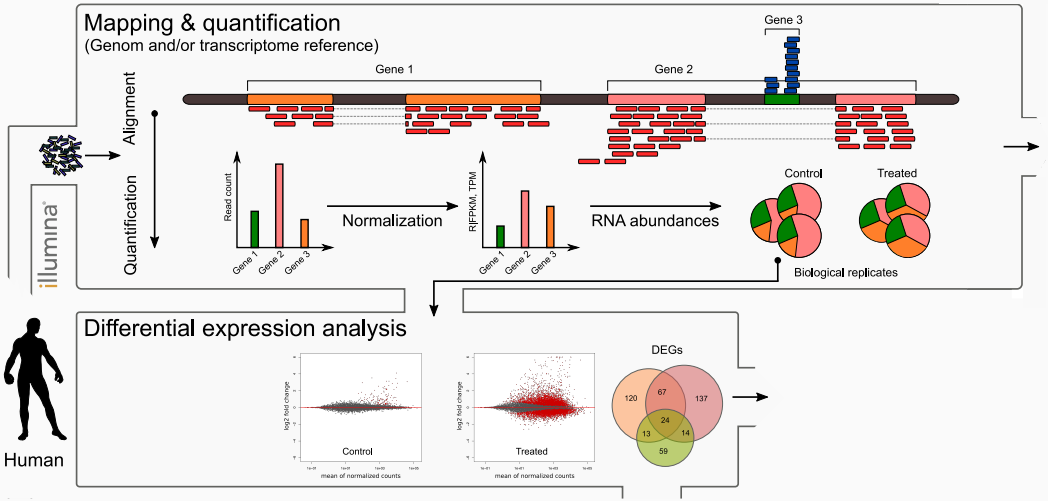
Reference-based RNA-Seq analysis



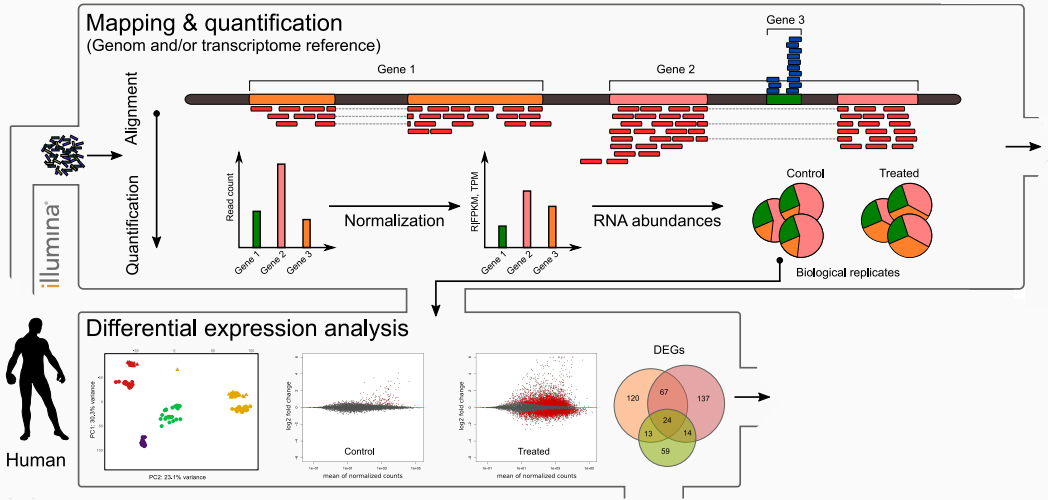
Human

```
hoelzer@prost:/mnt/prostlocal/projects/cellprogramming_fli_sarmistha/htseq/tophat$ head all.merged.htseq
ID      shSCR_iPSc_1    shSCR_iPSc_2    shSCR_iPSc_3    shTNFAIP2_iPSc_1        shTNFAIP2_iPSc_2        shTNFAIP2_iPSc_3
1       shTNFAIP2_noniPSc_2     shTNFAIP2_noniPSc_3
ENSMUSG00000000001      6000    5754    6116    5560    5083    4952    4865    4615    6246    4337    4231    3965
ENSMUSG00000000003      0       0       0       0       0       0       0       0       0       0       0       0
ENSMUSG00000000028      3282    3026    3147    2860    2760    2734    761     710     837     616     616     587
ENSMUSG00000000031      24789   21466   23701   17438   16474   16663   38593   40969   49883   29250   28070   25641
ENSMUSG00000000037      868     881     844     952     840     818     206     192     260     158     155     127
ENSMUSG00000000049      23      17      26      13      13      24      6       1       4       6       0       5
ENSMUSG00000000056      1875    1729    1832    1542    1432    1424    1304    1333    1685    1046    1007    909
ENSMUSG00000000058      155     155     144     168     124     159     1476    1315    1918    1635    1557    1592
ENSMUSG00000000078      2635    2527    2609    2482    2367    2206    4506    4495    5701    4538    4427    4053
```

Reference-based RNA-Seq analysis



Reference-based RNA-Seq analysis

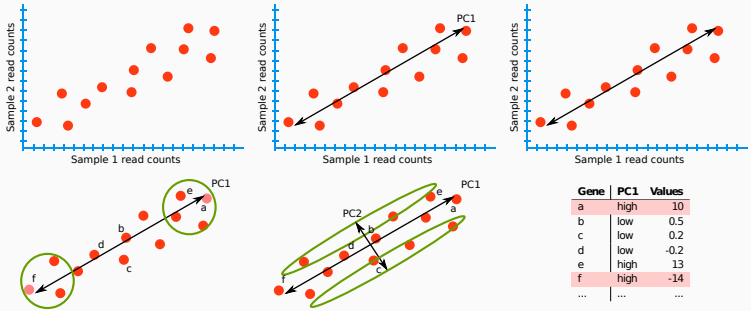


Input for PCA: RNA-Seq read counts

```
hoelzer@prost:/mnt/prostlocal/projects/cellprogramming_fli_sarmistha/htseq/tophat$ head all.merged.htseq
ID      shSCR_iPSc_1    shSCR_iPSc_2    shSCR_iPSc_3    shTNFAIP2_iPSc_1        shTNFAIP2_iPSc_2        shTNFAIP2_iPSc_3
1       shTNFAIP2_noniPSc_2     shTNFAIP2_noniPSc_3
ENSMUSG00000000001      6000    5754    6116    5560    5083    4952    4865    4615    6246    4337    4231    3965
ENSMUSG00000000003      0       0       0       0       0       0       0       0       0       0       0       0
ENSMUSG00000000028      3282    3026    3147    2860    2760    2734    761     710     837     616     616     587
ENSMUSG00000000031      24789   21466   23701   17438   16474   16663   38593   40969   49883   29250   28070   25641
ENSMUSG00000000037      868     881     844     952     840     818     206     192     260     158     155     127
ENSMUSG00000000049      23      17      26      13      13      24      6       1       4       6       0       5
ENSMUSG00000000056      1875    1729    1832    1542    1432    1424    1304    1333    1685    1046    1007    909
ENSMUSG00000000058      155     155     144     168     124     159     1476    1315    1918    1635    1557    1592
ENSMUSG00000000078      2635    2527    2609    2482    2367    2206    4506    4495    5701    4538    4427    4053
```

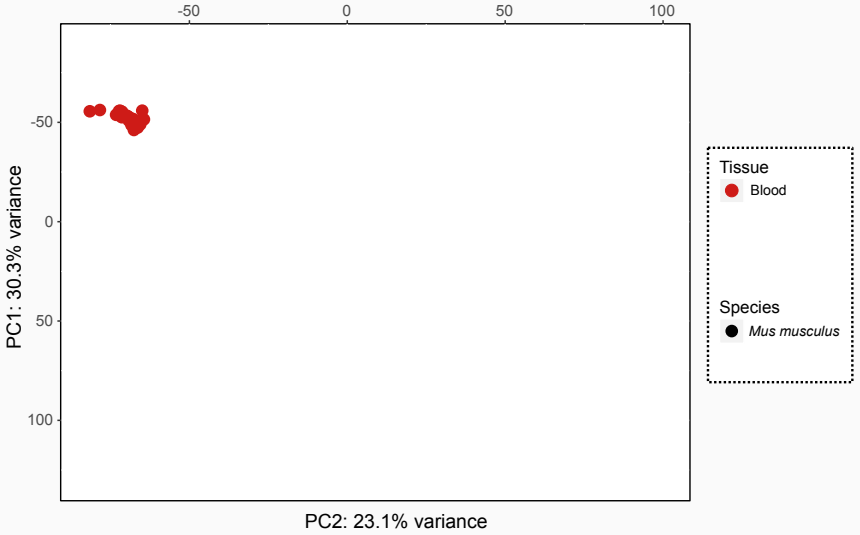
Input for PCA: RNA-Seq read counts

```
hoelzer@prost:/mnt/prostlocal/projects/cellprogramming_fli_sarmistha/htseq/tophat$ head all.merged.htseq
ID      shSCR_iPSc_1    shSCR_iPSc_2    shSCR_iPSc_3    shTNFAIP2_iPSc_1        shTNFAIP2_iPSc_2        shTNFAIP2_iPSc_3
1       shTNFAIP2_noniPSc_2     shTNFAIP2_noniPSc_3
ENSMUSG00000000001      6000    5754    6116    5560    5083    4952    4865    4615    6246    4337    4231    3965
ENSMUSG00000000003      0       0       0       0       0       0       0       0       0       0       0       0
ENSMUSG00000000028      3282    3026    3147    2860    2760    2734    761     710     837     616     616     587
ENSMUSG00000000031      24789   21466   23701   17438   16474   16663   38593   40969   49883   29250   28070   25641
ENSMUSG00000000037      868     881     844     952     840     818     206     192     260     158     155     127
ENSMUSG00000000049      23      17      26      13      13      24      6       1       4       6       0       5
ENSMUSG00000000056      1875    1729    1832    1542    1432    1424    1304    1333    1685    1046    1007    909
ENSMUSG00000000058      155     155     144     168     124     159     1476    1315    1918    1635    1557    1592
ENSMUSG00000000078      2635    2527    2609    2482    2367    2206    4506    4495    5701    4538    4427    4053
```

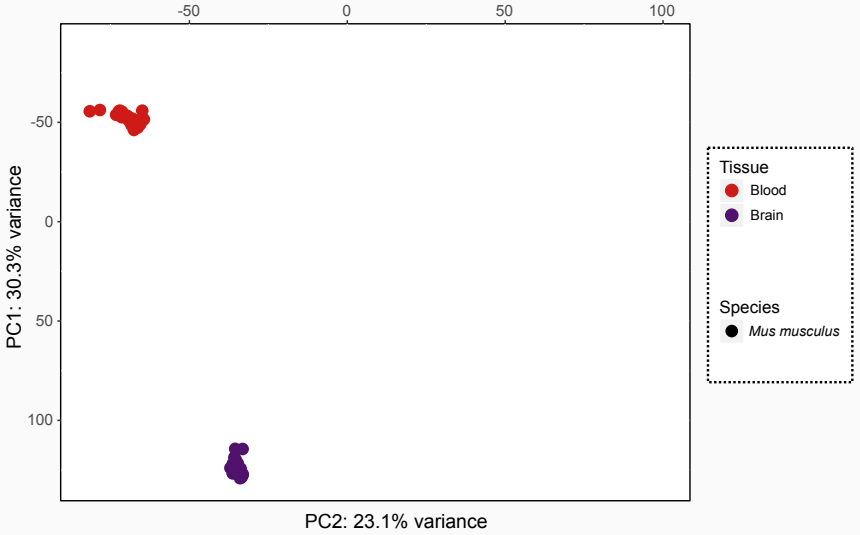


Gene	PC1	Values
a	high	10
b	low	0.5
c	low	0.2
d	low	-0.2
e	high	13
f	high	-14
...	...	...

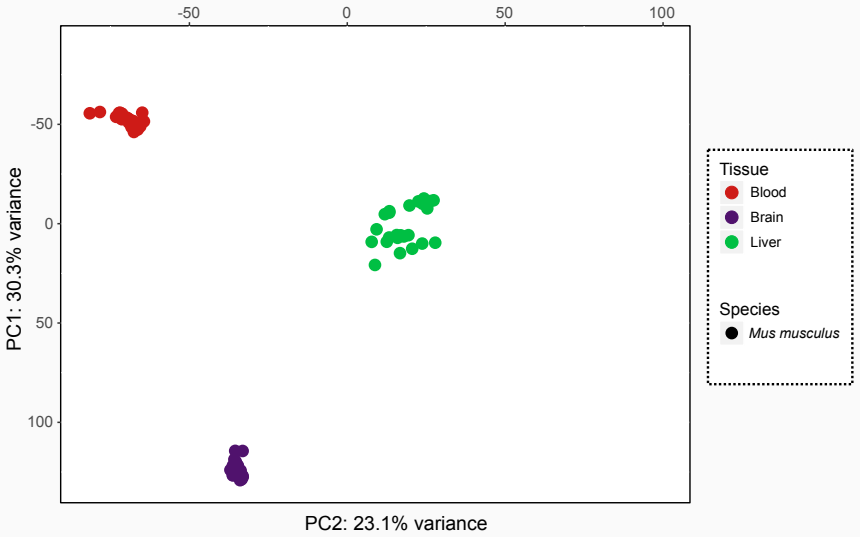
RNA-Seq PCA example, two dimensions



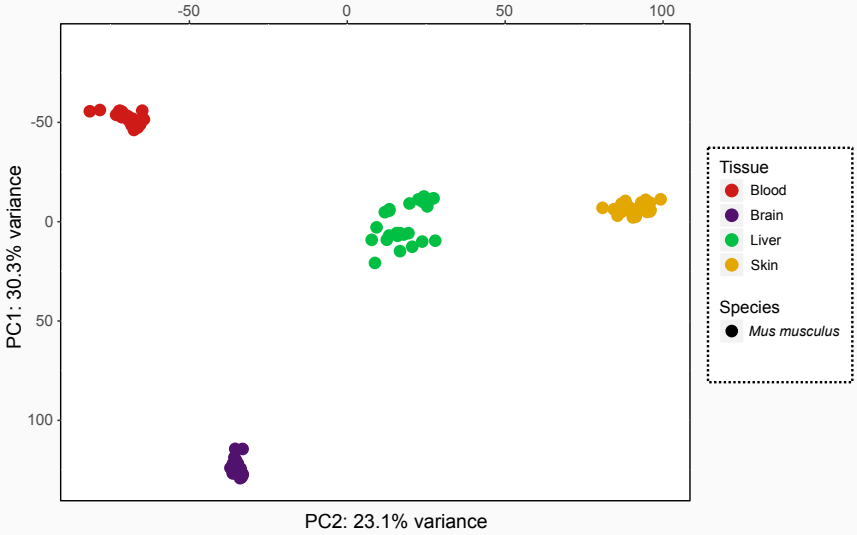
RNA-Seq PCA example, two dimensions



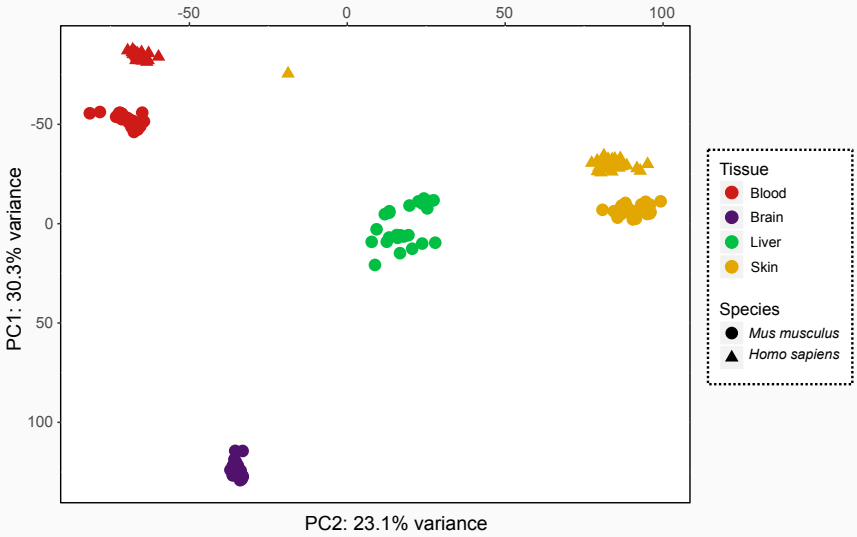
RNA-Seq PCA example, two dimensions



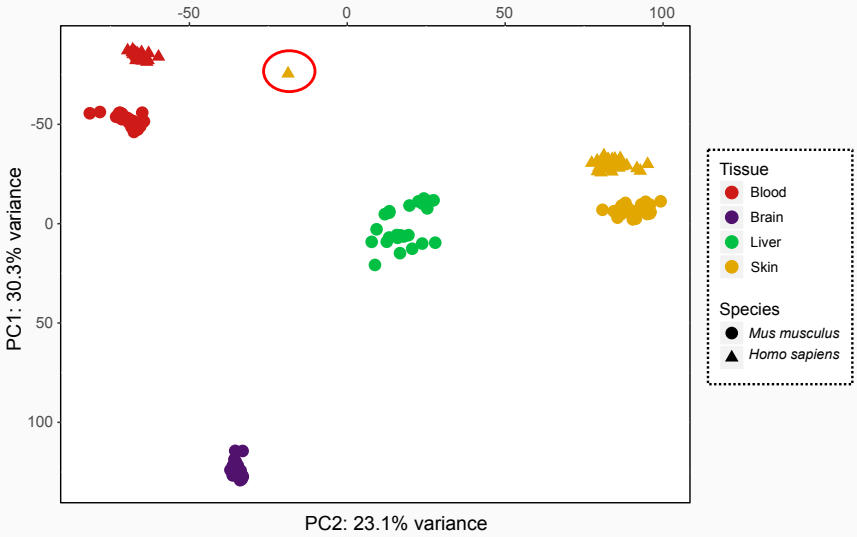
RNA-Seq PCA example, two dimensions



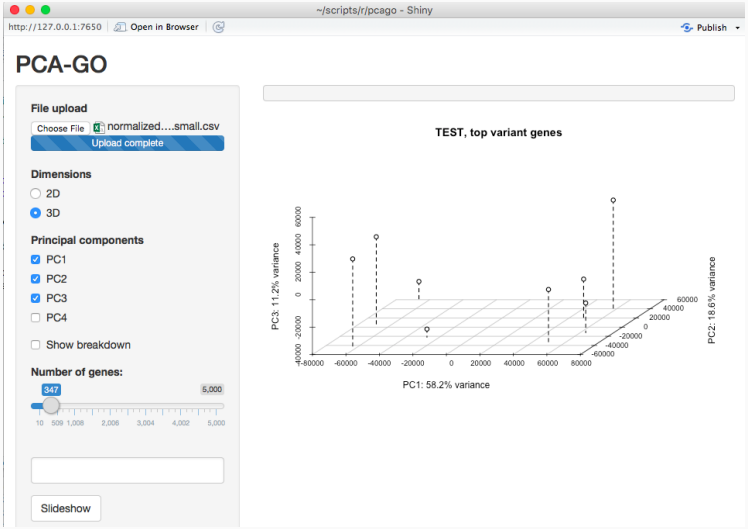
RNA-Seq PCA example, two dimensions



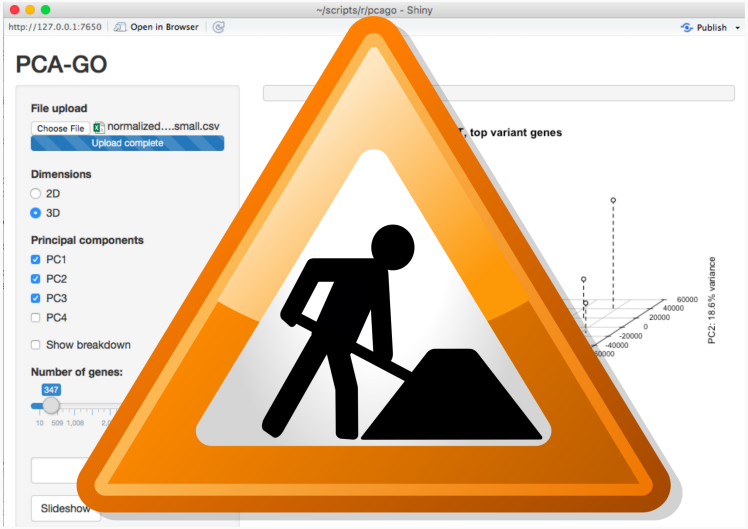
RNA-Seq PCA example, two dimensions



2016: Would be nice to have an interactive PCA tool...



2016: Would be nice to have an interactive PCA tool...



2018: Tadaa!

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." Manuscript in preparation. Web service already available at `http://pcago.bioinf.uni-jena.de/`

Hands on PCAGO

Outlook

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...

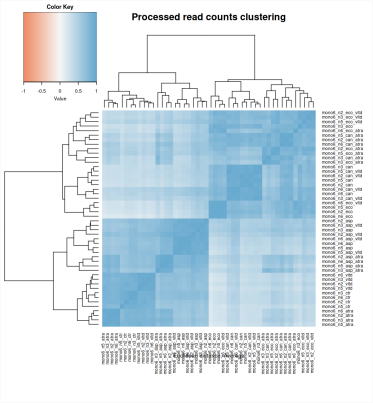
- `pcago.bioinf.uni-jena.de`
- `https://github.com/rumangerst/pcago`

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...

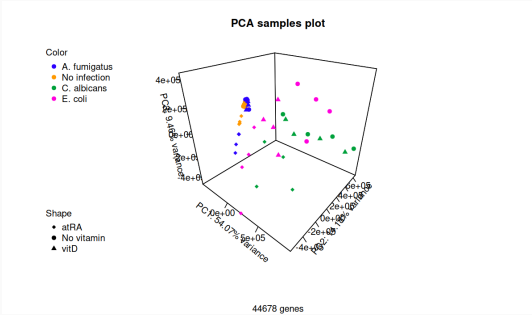
- `pcago.bioinf.uni-jena.de`
- `https://github.com/rumangerst/pcago`



Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...

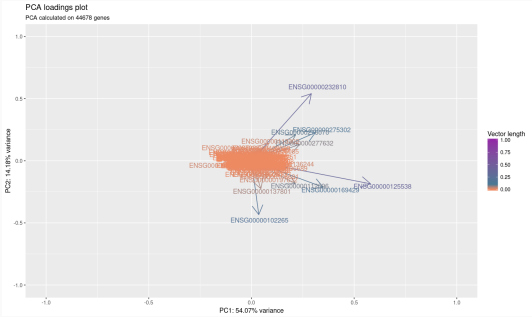


- pcago.bioinf.uni-jena.de
- https://github.com/rumangerst/pcago

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...

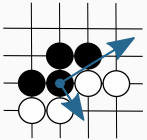


- `pcago.bioinf.uni-jena.de`
- `https://github.com/rumangerst/pcago`

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...



PCAGO

- pcago.bioinf.uni-jena.de
- https://github.com/rumangerst/pcago

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

Outlook

- improve usability
- include more features:
  - correlation heat maps
  - perspective 3D-plots (rotatable)
  - PCA loading plots
    (impact of single genes)
  - ...

- `pcago.bioinf.uni-jena.de`
- `https://github.com/rumangerst/pcago`

Ruman Gerst, Manja Marz and Martin Hölzer. "PCAGO: An interactive web service for analyzing RNA-Seq data with principal component analysis." In preparation.

... and all co-operation partners!