Визуализация сигнальных путей по данным экспрессии генов

Студенты:
Ольга Романова
Станислав Легковой

Руководители:
Оксана Станевич (ПСПбГМУ им. акад. И.П.Павлова)
Евгений Бакин (Центр геномной биоинформатики им. Ф.Г. Добржанского)

Цель и задачи проекта

Цель - визуализация сигнальных путей по данным экспрессии генов, полученных из баз данных NCBI

Задачи:

- изучение понятий «сигнальный путь» и «дифференциальная экспрессия генов»
- освоение web-инструмента Phantasus для анализа экспрессии генов из конкретного датасета
- освоение пакетов pathview, gage и GEOquery в R для визуализации сигнальных путей с помощью графа
- реализация алгоритмов в R, позволяющих напрямую получать данные экспрессии генов из баз данных NCBI с целью поиска наиболее значимых сигнальных путей из KEGG PATHWAY

Цель и задачи проекта

Цель - визуализация сигнальных путей по данным экспрессии генов, полученных из баз данных NCBI

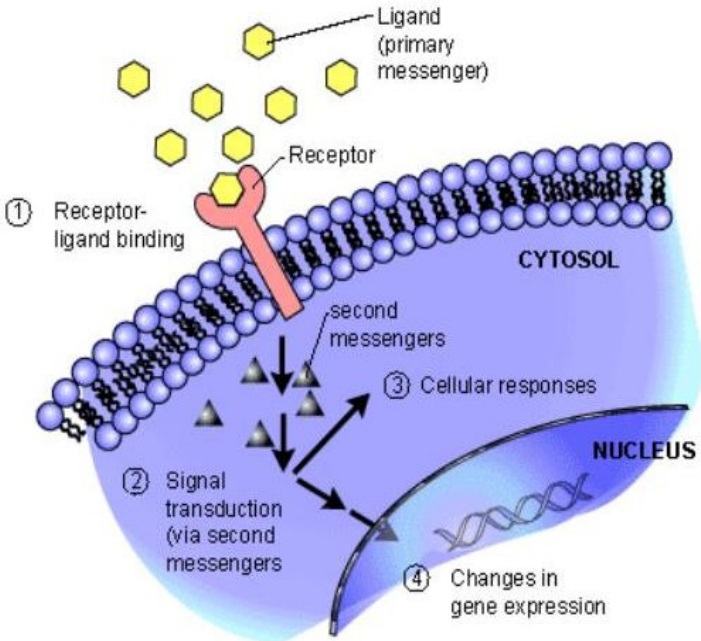
Задачи:

- ✓ изучение понятий «сигнальный путь» и «дифференциальная экспрессия генов»
- ➢ освоение web-инструмента Phantasus для анализа экспрессии генов из конкретного датасета
- ➢ освоение пакетов pathview, gage и GEOquery в R для визуализации сигнальных путей с помощью графа
- ➢ реализация алгоритмов в R, позволяющих напрямую получать данные экспрессии генов из баз данных NCBI с целью поиска наиболее значимых сигнальных путей из KEGG PATHWAY

Дифференциальная экспрессия генов

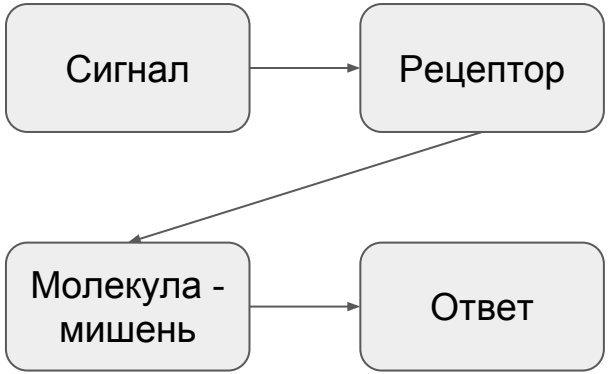
Геном организма - это полный набор генов в каждом из его клеток. Учитывая особенности организма, каждая клетка имеет точную копию одного и того же генома, однако:

- разные клетки экспрессируют разные гены
- разные гены экспрессируются в разных условиях
- дифференциальная экспрессия генов ведет к измененному состоянию клетки



Сигнальный путь

совокупность молекулярных изменений, преобразующих сигнал на поверхности клетки в ответ внутри клетки



Цель и задачи проекта

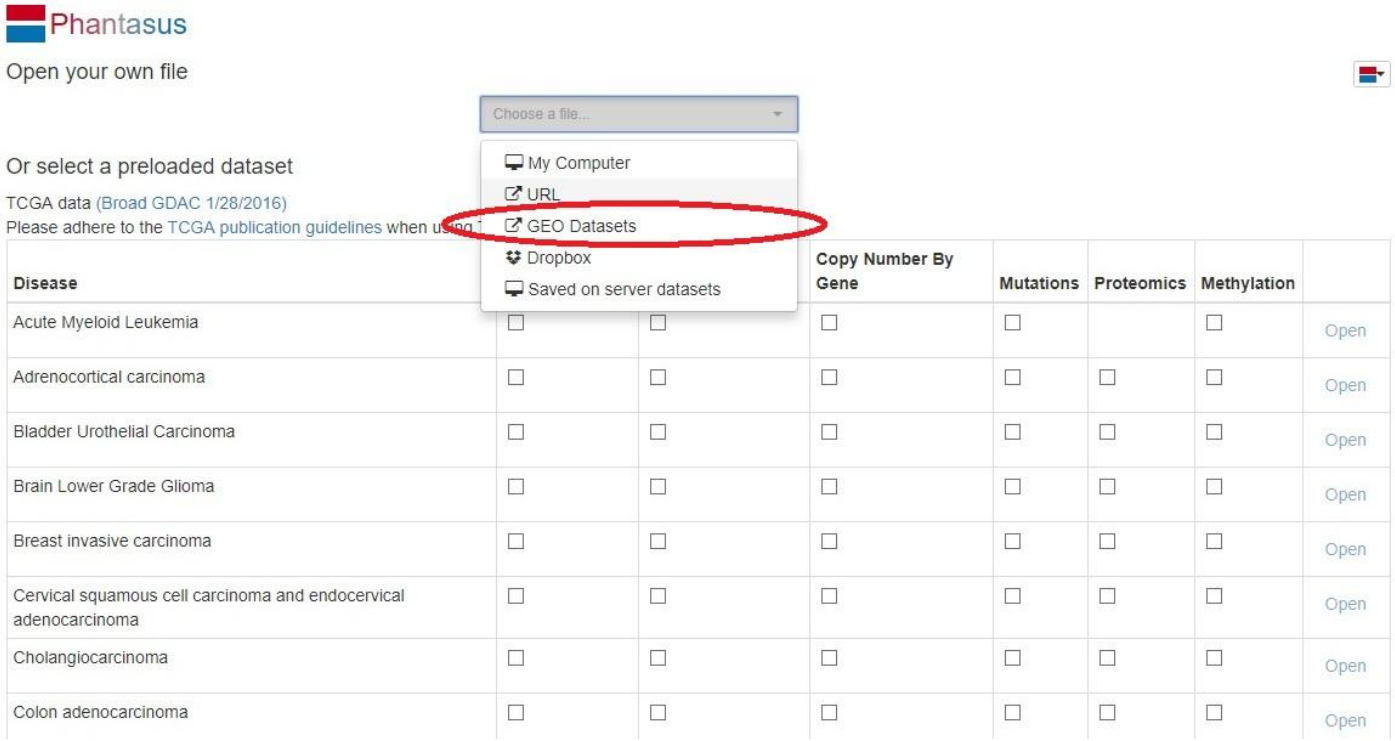
Цель - визуализация сигнальных путей по данным экспрессии генов, полученных из баз данных NCBI

Задачи:

- ✓ изучение понятий «сигнальный путь» и «дифференциальная экспрессия генов»
- ✓ освоение web-инструмента Phantasus для анализа экспрессии генов из конкретного датасета
- ➢ освоение пакетов pathview, gage и GEOquery в R для визуализации сигнальных путей с помощью графа
- ➢ реализация алгоритмов в R, позволяющих напрямую получать данные экспрессии генов из баз данных NCBI с целью поиска наиболее значимых сигнальных путей из KEGG PATHWAY

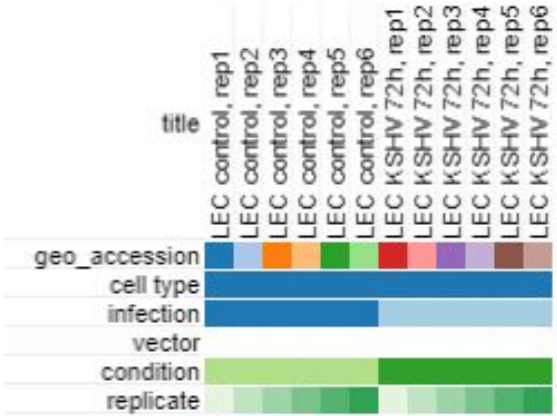
Phantasus

веб-приложение для исследовательского анализа экспрессии генов с поддержкой R
(создатели: Дарья Зенкова, Владислав Каменев, Максим Артёмов, Алексей Сергушичев)



GSE16357-GPL570 × GSE16357-GPL8617 × GSE16357-GPL570 ×

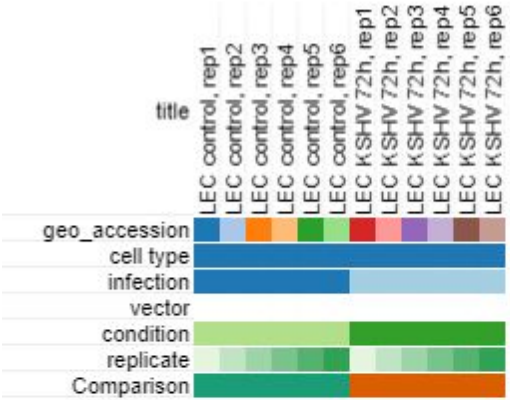
File Edit View Tools Help Rows Columns 54,675 rows by 12 columns



id	Gene symbol	Gene ID
1007_s_at	MIR4640///DDR1	100616237///780
1053_at	RFC2	5982
117_at	HSPA6	3310
121_at	PAX8	7849
1255_g_at	GUCA1A	2978
1294_at	MIR5193///UBA7	100847079///7318
1316_at	THRA	7067
1320_at	PTPN21	11099
1405_i_at	CCL5	6352
1431_at	CYP2E1	1571
1438_at	EPHB3	2049
1487_at	ESRRA	2101
1494_f_at	CYP2A6	1548
1552256_a_at	SCARB1	949
1552257_a_at	TTLL12	23170
1552258_at	LOC101930489///LINC00152	101930489///112597
1552261_at	WFDC2	10406
1552263_at	MAPK1	5594

GSE16357-GPL570 × GSE16357-GPL8617 × GSE16357-GPL570 × GSE16357-GPL570 ×

File Edit View Tools Help Rows Columns 22,190 rows by 12 columns 0 r



Gene ID	mean_val	logFC	AveExpr	t	P.Value	adj.P.Val ▲
4312	8.8051	8.6748	8.8051	139.27	1.7513e-25	3.8861e-21
81606	8.2626	6.2549	8.2626	83.481	4.9211e-22	5.4599e-18
81848	9.0626	4.0592	9.0626	69.628	8.1902e-21	5.0743e-17
5367	8.8460	6.2385	8.8460	69.133	9.1470e-21	5.0743e-17
5999	9.4665	-4.1225	9.4665	-60.404	7.3939e-20	3.2814e-16
11080	9.9729	-2.3538	9.9729	-58.324	1.2717e-19	4.7033e-16
6781	9.6627	3.6783	9.6627	57.262	1.6897e-19	5.3564e-16
7076	12.780	2.4093	12.780	56.251	2.2256e-19	6.1732e-16
760	10.259	4.8026	10.259	55.401	2.8168e-19	6.9449e-16
2321	8.2351	3.4340	8.2351	53.315	5.0973e-19	1.1311e-15
1604	10.897	3.2972	10.897	50.544	1.1629e-18	2.3458e-15
3142	8.1603	3.8863	8.1603	49.318	1.6996e-18	3.1429e-15
3759	7.8814	5.3638	7.8814	48.611	2.1241e-18	3.6256e-15
8644	8.0200	-3.6111	8.0200	-45.125	6.6980e-18	1.0616e-14
4860	11.880	2.2361	11.880	44.850	7.3592e-18	1.0887e-14
55859	8.4260	-2.5704	8.4260	-43.781	1.0676e-17	1.3936e-14
1520	7.0779	2.4851	7.0779	43.904	1.0224e-17	1.3936e-14
64399	8.6650	-2.7908	8.6650	-42.977	1.4208e-17	1.7515e-14
57088	9.3872	-4.0028	9.3872	-42.713	1.5624e-17	1.8248e-14
5010	6.3178	-2.3081	6.3178	-42.327	1.7968e-17	1.9936e-14
359845	9.8604	2.9276	9.8604	41.821	2.1631e-17	2.2856e-14
161742	8.2102	2.5370	8.2102	40.699	3.2881e-17	3.3165e-14

Цель и задачи проекта

Цель - визуализация сигнальных путей по данным экспрессии генов, полученных из баз данных NCBI

Задачи:

- ✓ изучение понятий «сигнальный путь» и «дифференциальная экспрессия генов»
- ✓ освоение web-инструмента Phantasus для анализа экспрессии генов из конкретного датасета
- ✓ освоение пакетов pathview, gage и GEOquery в R для визуализации сигнальных путей с помощью графа
- ➢ реализация алгоритмов в R, позволяющих напрямую получать данные экспрессии генов из баз данных NCBI с целью поиска наиболее значимых сигнальных путей из KEGG PATHWAY

KEGG PATHWAY (Kyoto Encyclopedia of Genes and Genomes Pathway)

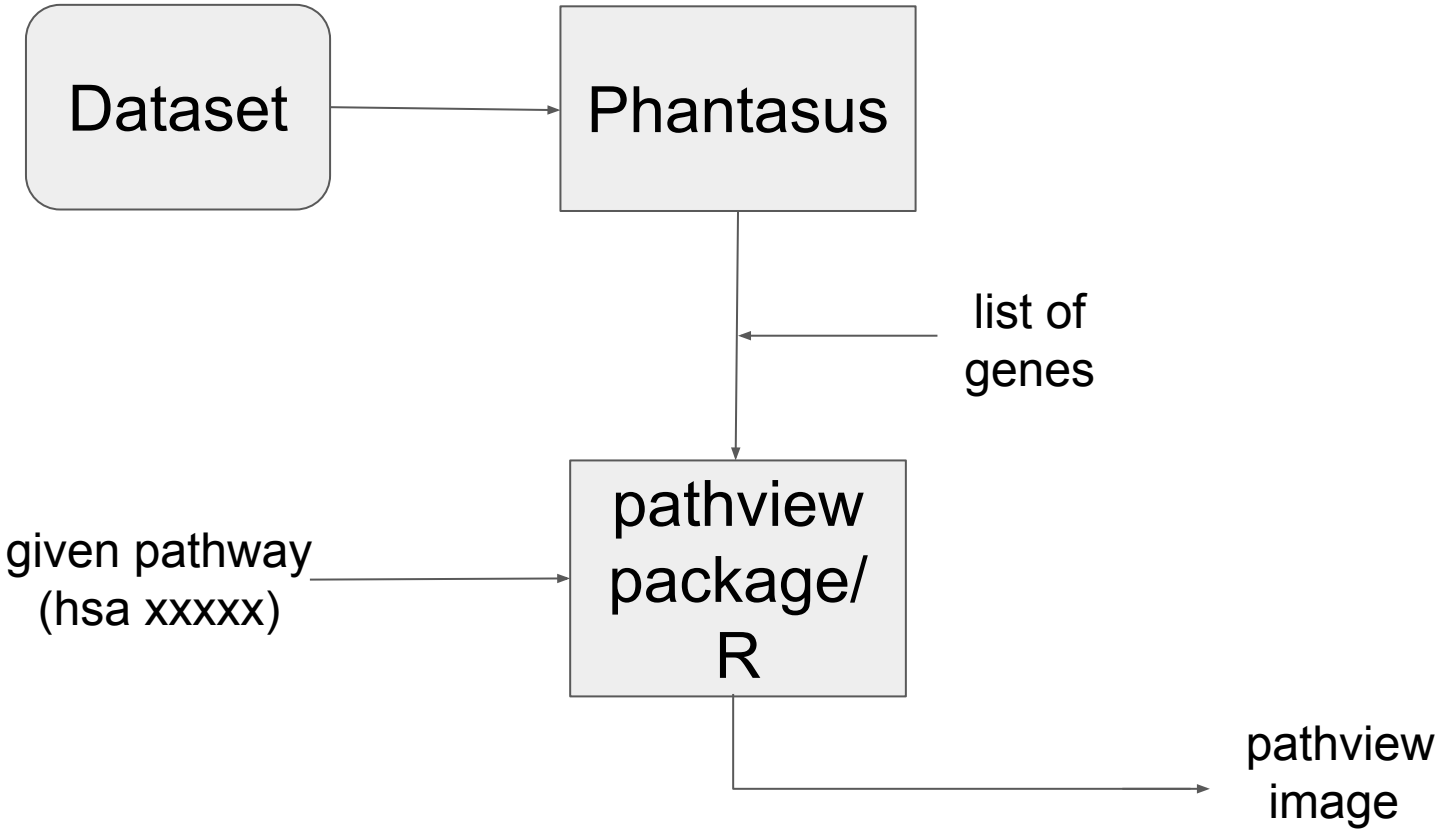
База данных, содержащая информацию о генных продуктах, связанных в сети белок-белковых взаимодействий, сети ферментов и регуляторов генов; данные представлены в виде графических диаграмм, включающих большинство метаболических путей и некоторые из наиболее известных регуляторных путей.

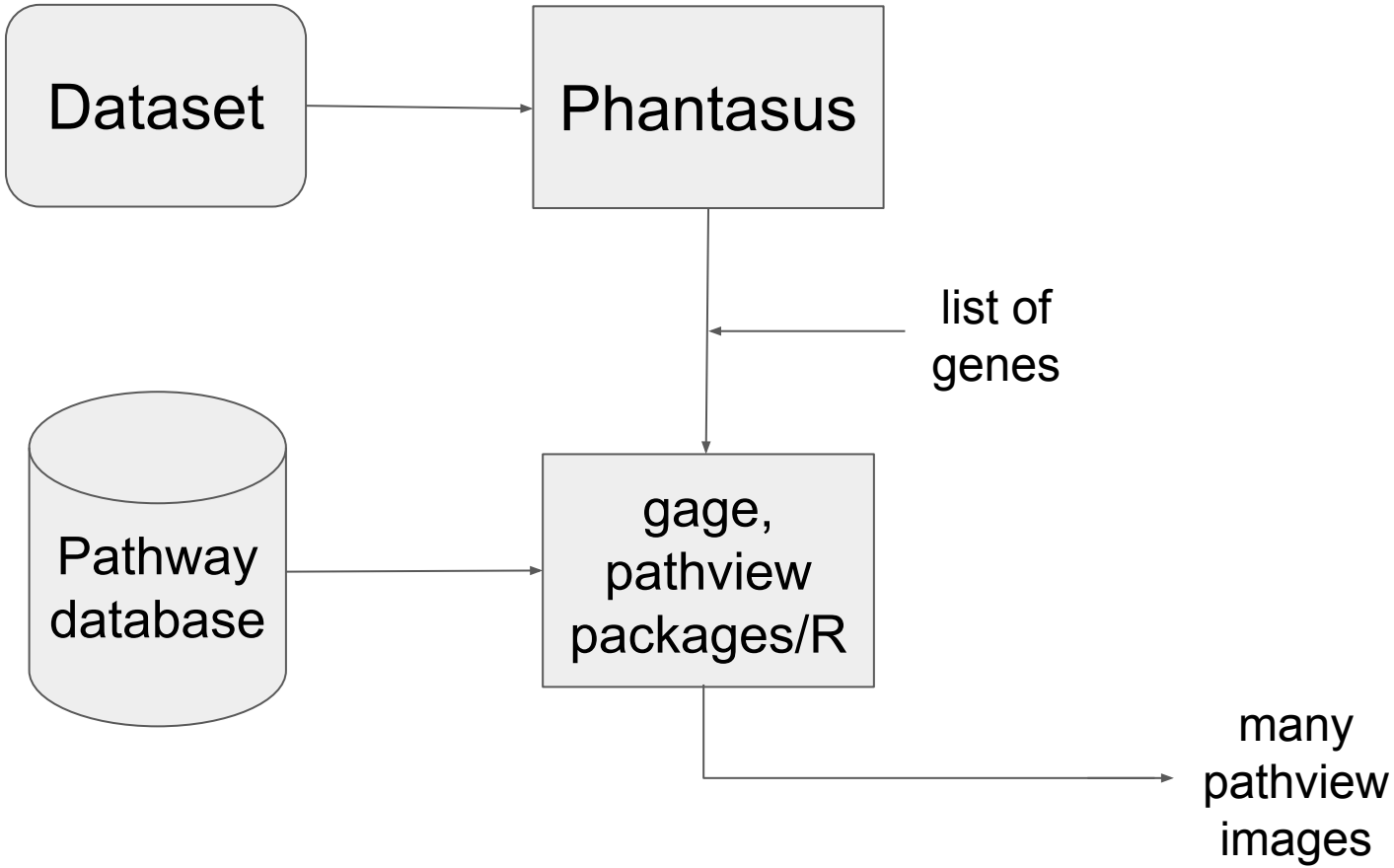
Цель и задачи проекта

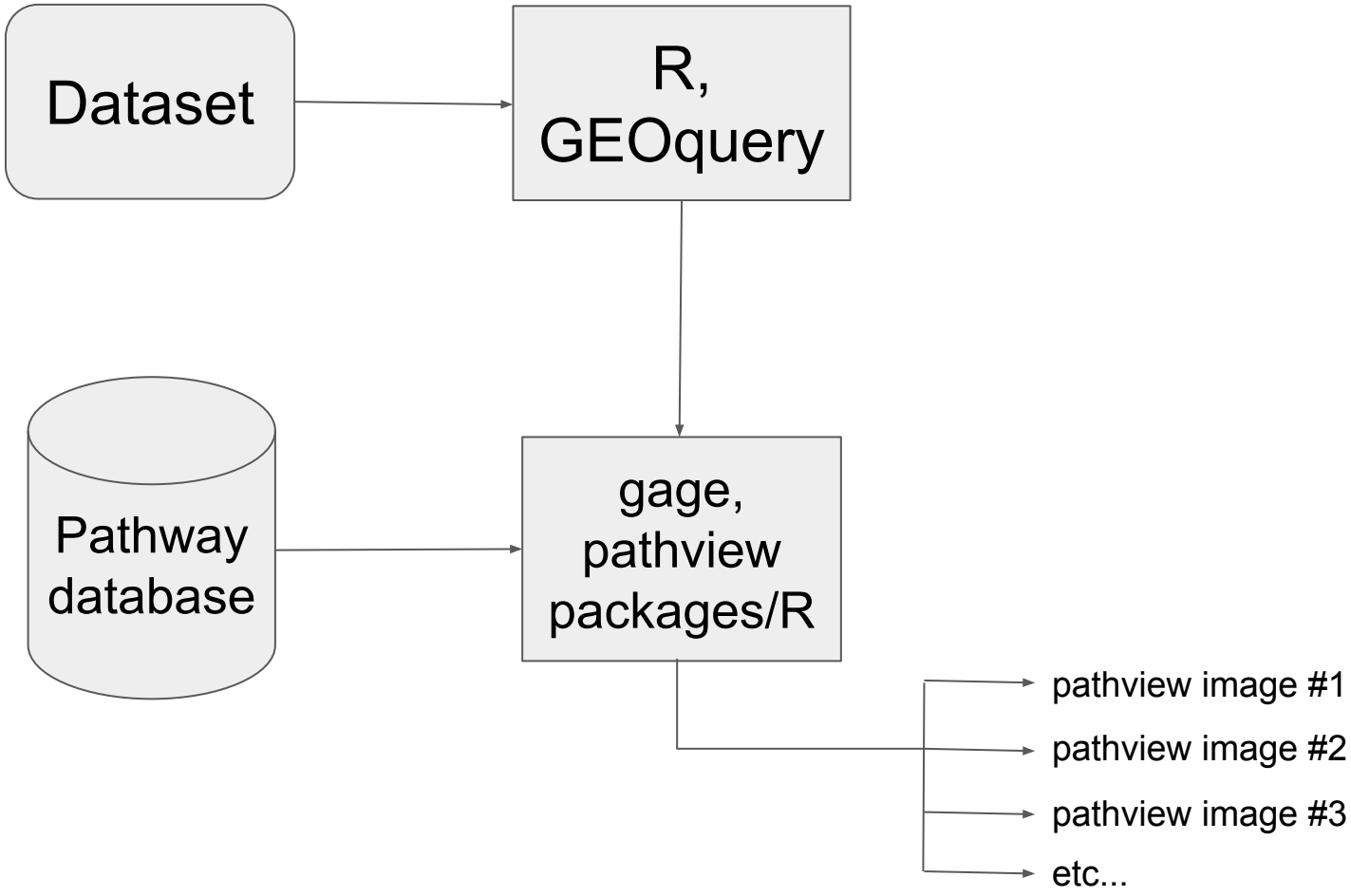
Цель - визуализация сигнальных путей по данным экспрессии генов, полученных из баз данных NCBI

Задачи:

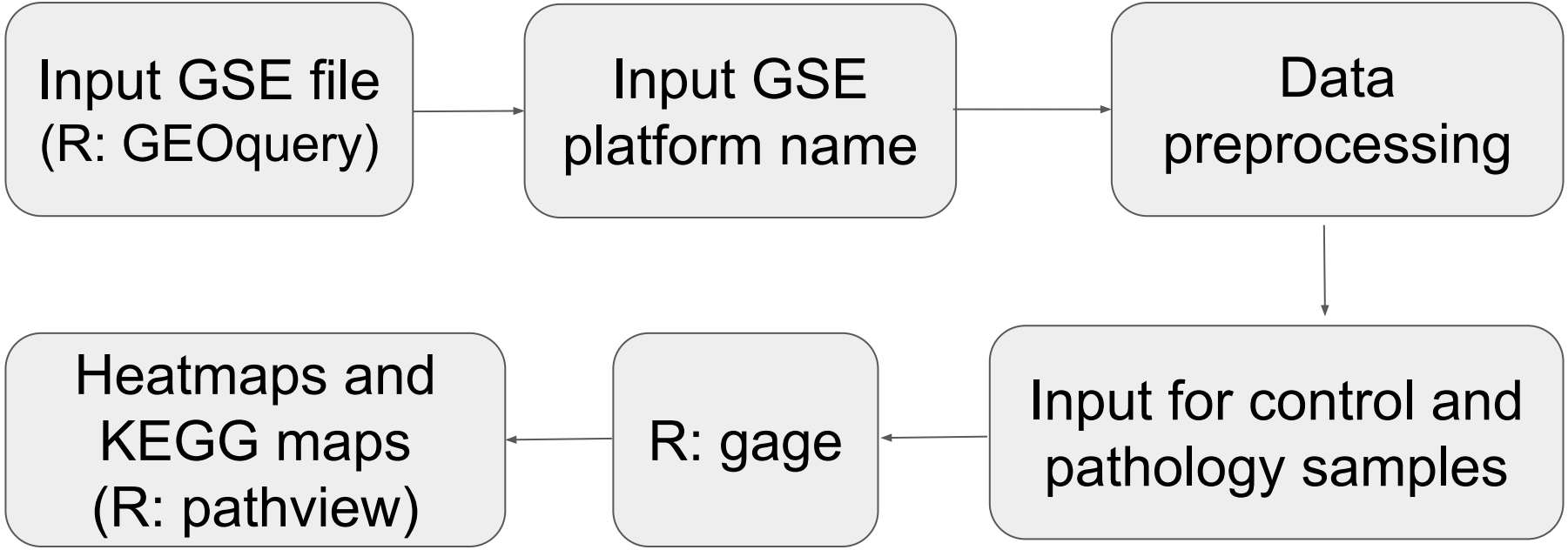
- ✓ изучение понятий «сигнальный путь» и «дифференциальная экспрессия генов»
- ✓ освоение web-инструмента Phantasus для анализа экспрессии генов из конкретного датасета
- ✓ освоение пакетов pathview, gage и GEOquery в R для визуализации сигнальных путей с помощью графа
- ✓ реализация алгоритмов в R, позволяющих напрямую получать данные экспрессии генов из баз данных NCBI с целью поиска наиболее значимых сигнальных путей из KEGG PATHWAY


Dataset
Phantasus
list of genes
given pathway (hsa xxxxx)
pathview package/ R
pathview image


Dataset
Phantasus
list of genes
Pathway database
gage, pathview packages/R
many pathview images


Dataset
R,
GEOquery
Pathway
database
gage,
pathview
packages/R
pathview image #1
pathview image #2
pathview image #3
etc...

Алгоритм в R - визуализация сигнальных путей



Console Terminal

D:/Project/result/

```
> library(data.table)
> library(pathview)
> library(gage)
> library("org.Hs.eg.db")
> library(clusterProfiler)
> library(dplyr)
> gse_name <- readline(prompt = "Enter GSE file name:")
Enter GSE file name:GSE16357
```

Console Terminal

D:/Project/result/

```
> data_gse16357 <- setDT(as.data.frame(data_gse16357), keep.rownames = TRUE)[]
> data_gse16357 <- data_gse16357 %>%
+   group_by(rn) %>%
+   summarise_all(mean)
> data_gse16357_temp <- data_gse16357
> data_gse16357 <- data_gse16357[-c(1)]
> rownames(data_gse16357) <- data_gse16357_temp$rn
Warning message:
Setting row names on a tibble is deprecated.
> rm(data_gse16357_temp)
> colnames(data_gse16357)
 [1] "GSM410666" "GSM410667" "GSM410668" "GSM410669" "GSM410670" "GSM410671" "GSM410672" "GSM410673"
 [9] "GSM410674" "GSM410675" "GSM410676" "GSM410677"
> con_sam <- readline(prompt = "Enter numbers of control samples: ")
Enter numbers of control samples: 1,2,3,
```

Console Terminal

D:/Browser_downloads/

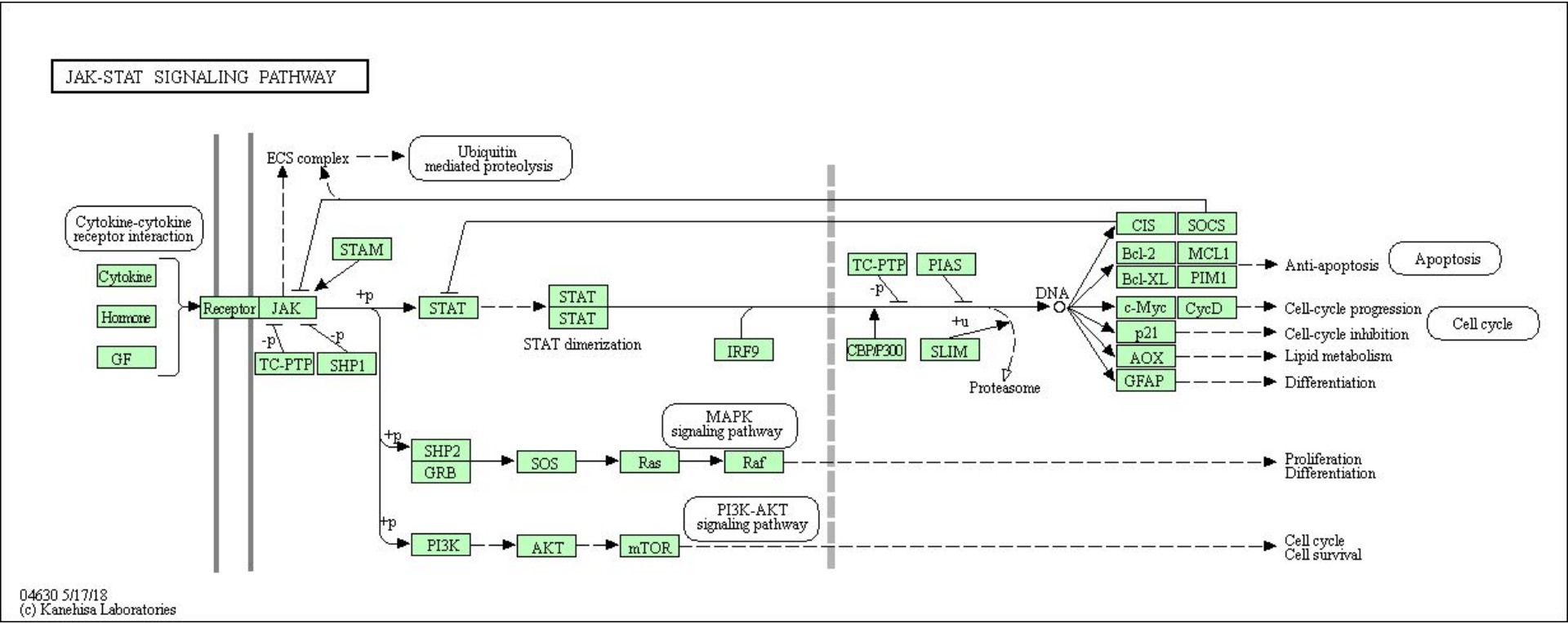
```
> library(clusterProfiler)
clusterProfiler v3.8.1  For help: https://guangchuangyu.github.io/software/clusterProfiler

If you use clusterProfiler in published research, please cite:
Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R package for comparing biological the
mes among gene clusters. OMICS: A Journal of Integrative Biology. 2012, 16(5):284-287.
> exprss_data.kegg.p <- gage(data_gse16357, gsets = kegg.gs, ref = control_samples, samp = disease_samples)
> exprss_data.kegg.sig<-sigGeneSet(exprss_data.kegg.p, outname="exprss_data.kegg")
[1]  "there are 50 signficantly up-regulated gene sets"
[1] "there are 37 signficantly down-regulated gene sets"
>
```

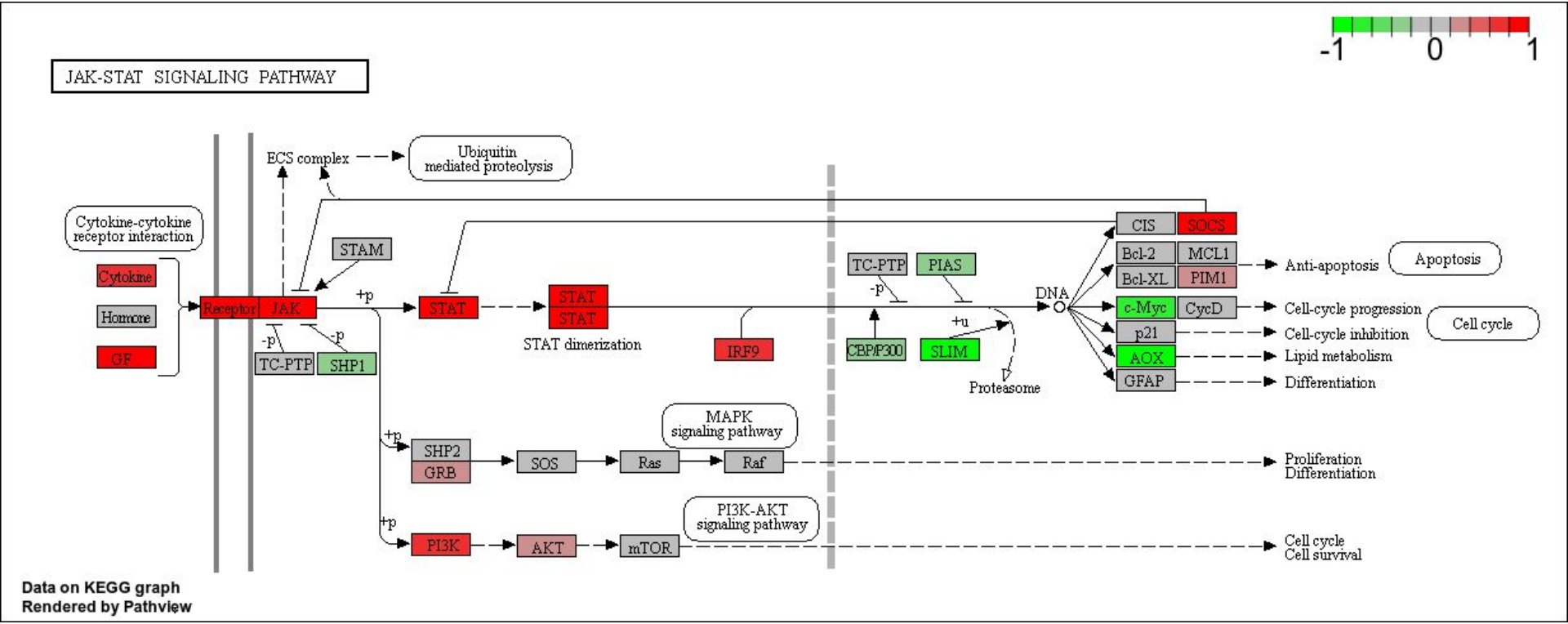
Console ~/

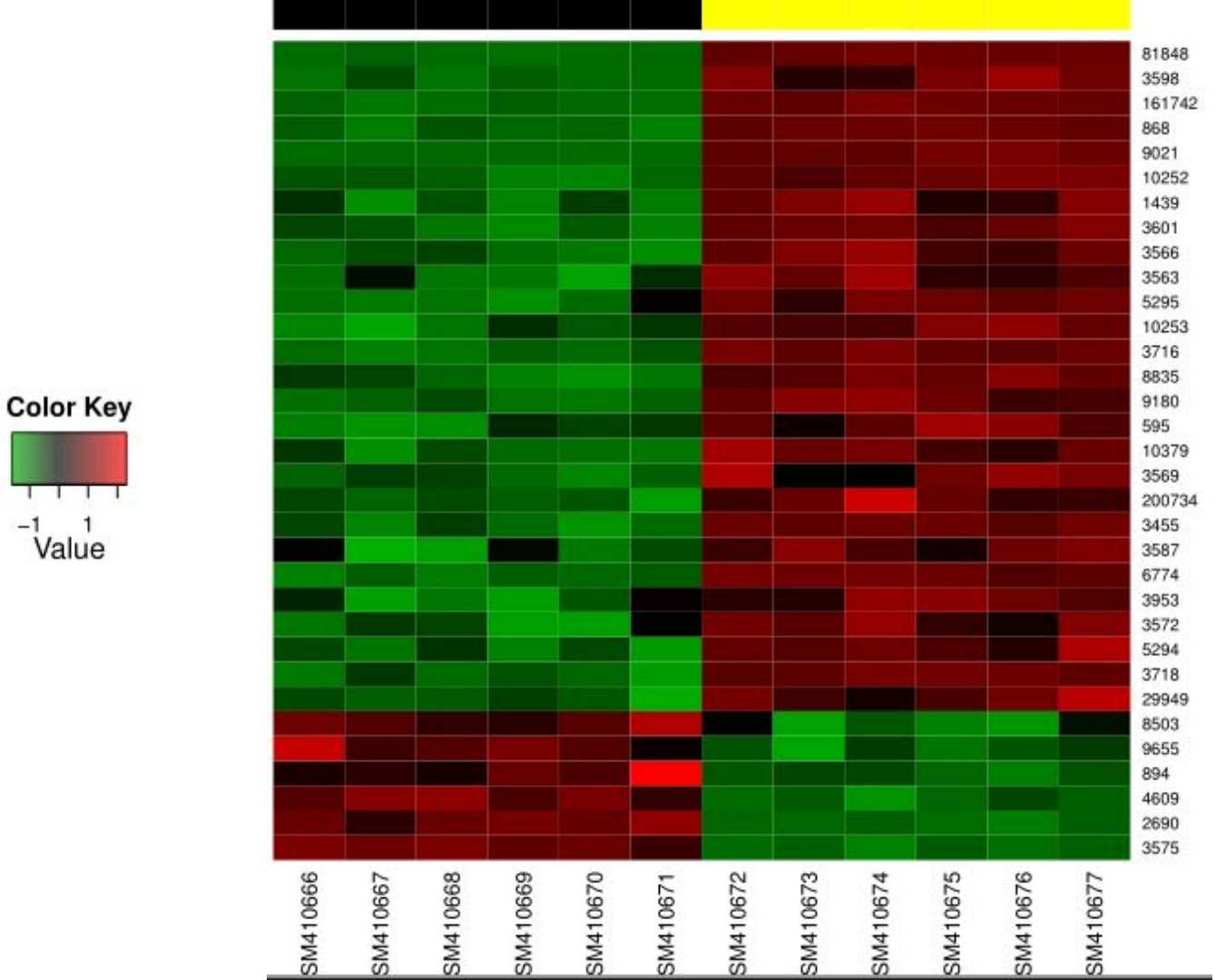
```
Info: Downloading png files for hsa03010, 1/1 pathways..
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /home/legosta
Info: Writing image file hsa03010.pathview.png
Info: Downloading xml files for hsa04110, 1/1 pathways..
Info: Downloading png files for hsa04110, 1/1 pathways..
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /home/legosta
Info: Writing image file hsa04110.pathview.png
Info: Downloading xml files for hsa03013, 1/1 pathways..
Info: Downloading png files for hsa03013, 1/1 pathways..
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /home/legosta
Info: Writing image file hsa03013.pathview.png
```

Сигнальный путь JAK-STAT

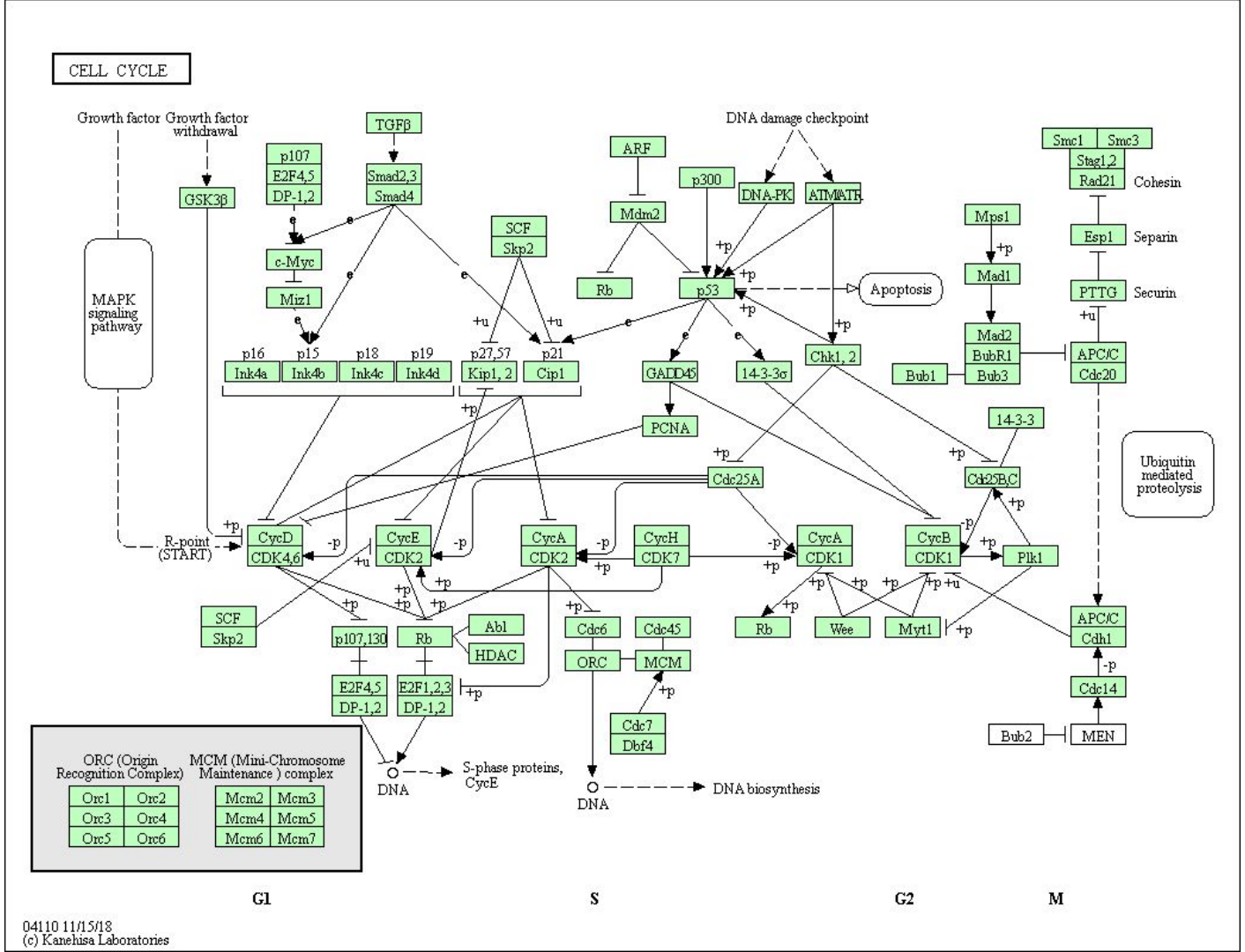


Сигнальный путь JAK-STAT

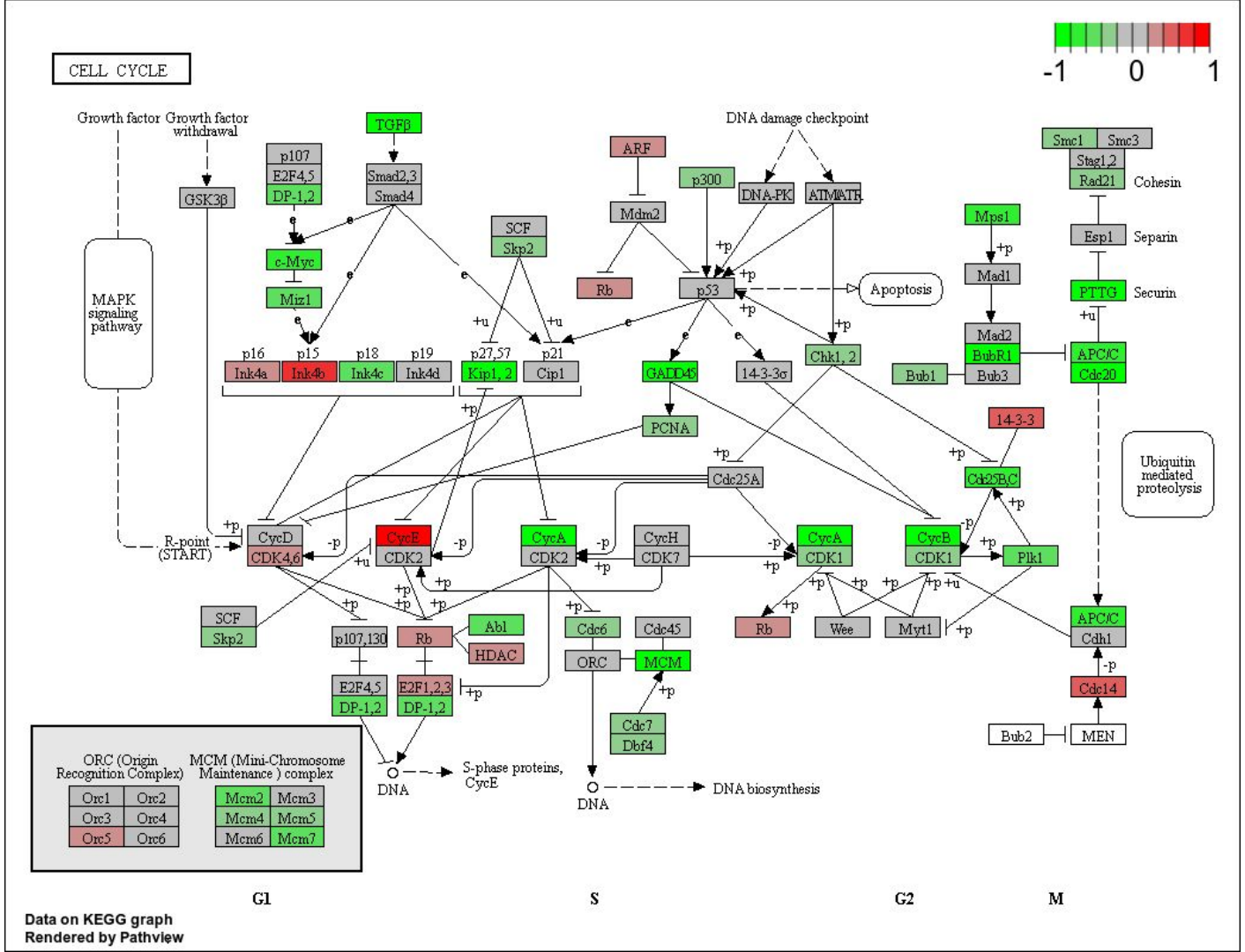


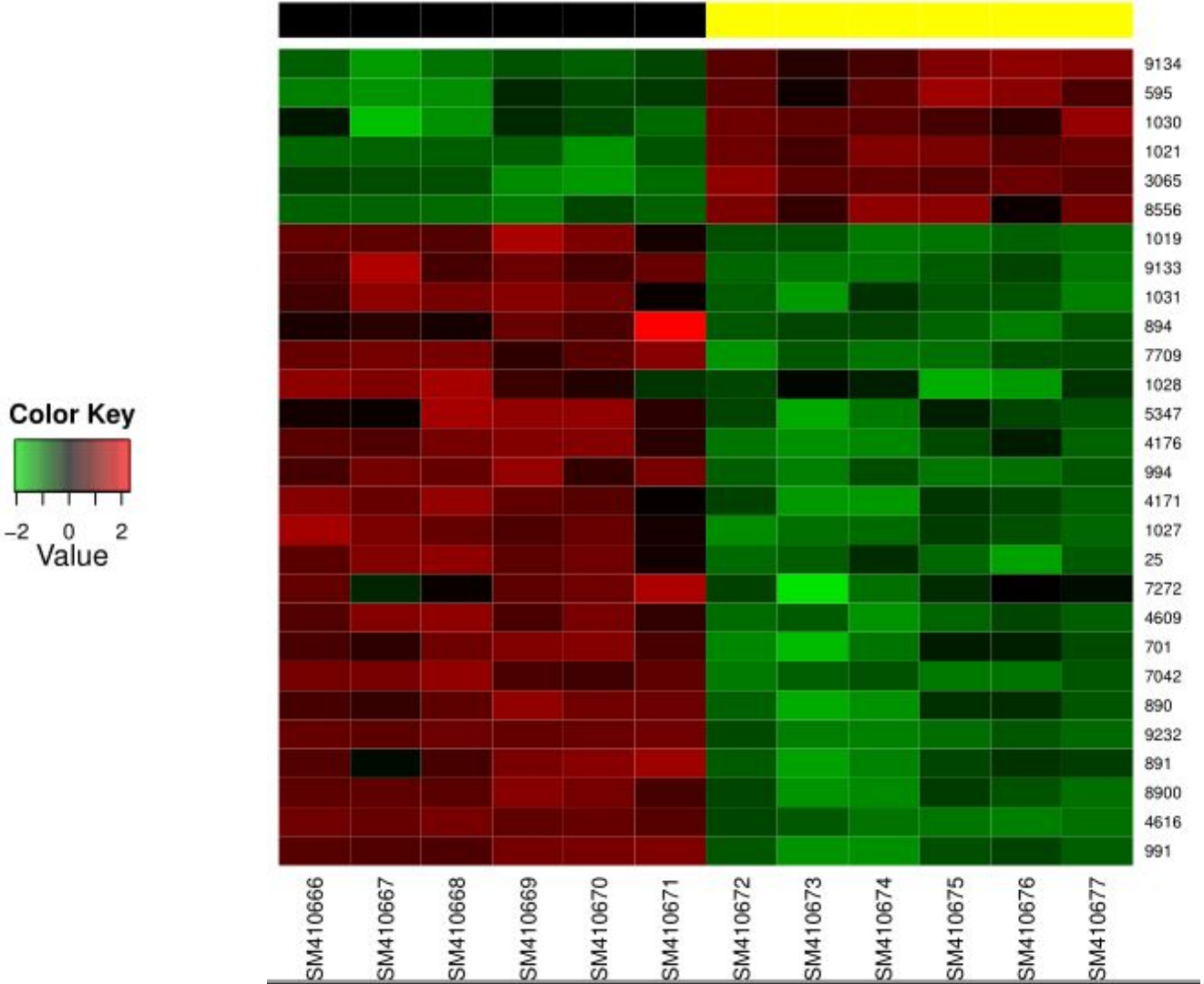

Color Key
−1
1
Value
81848
3598
161742
868
9021
10252
1439
3601
3566
3563
5295
10253
3716
8835
9180
595
10379
3569
200734
3455
3587
6774
3953
3572
5294
3718
29949
8503
9655
894
4609
2690
3575
SM410666
SM410667
SM410668
SM410669
SM410670
SM410671
SM410672
SM410673
SM410674
SM410675
SM410676
SM410677

Клеточный цикл клетки



Клеточный цикл клетки



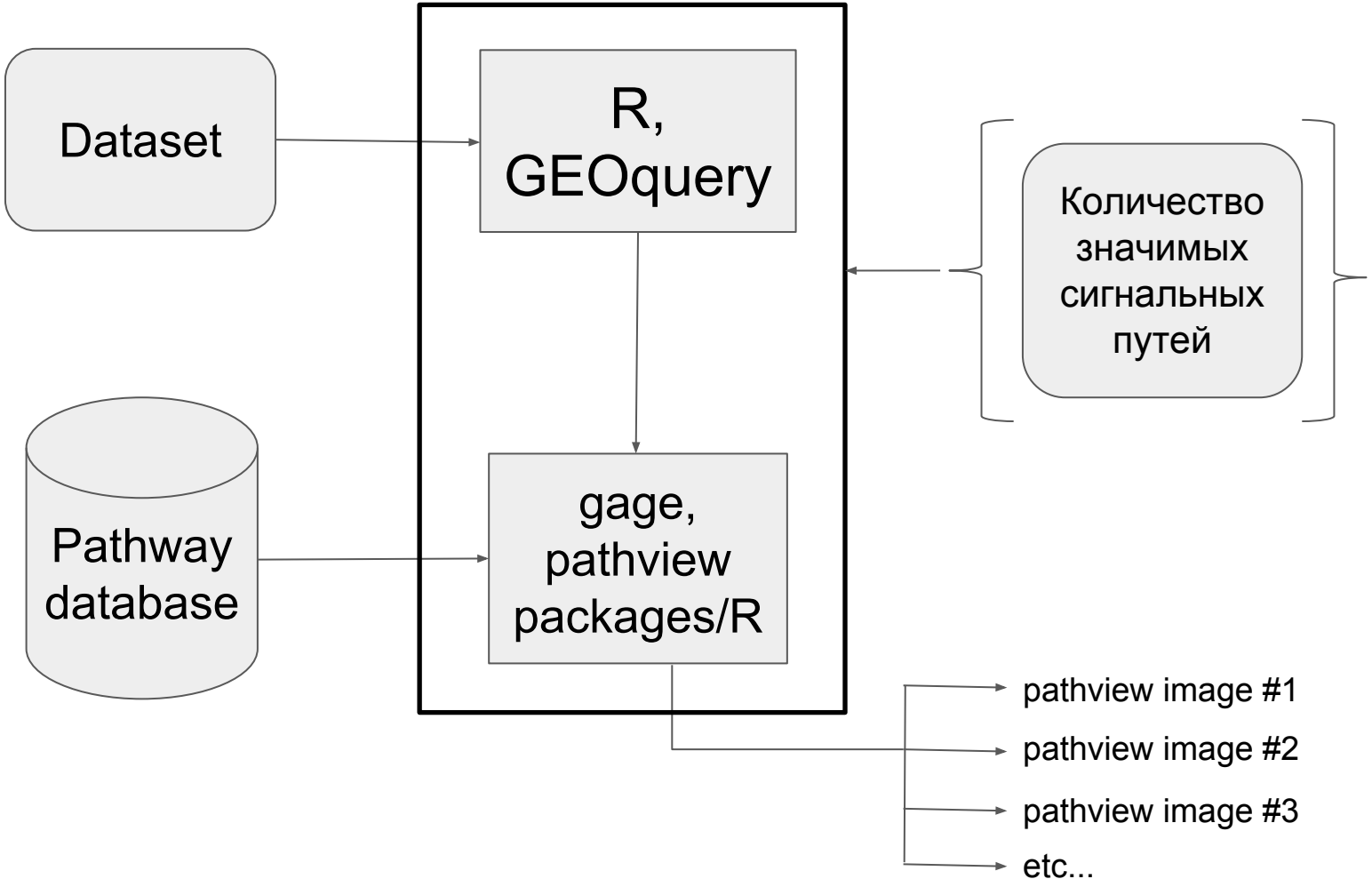

Color Key
−2
0
2
Value
9134
595
1030
1021
3065
8556
1019
9133
1031
894
7709
1028
5347
4176
994
4171
1027
25
7272
4609
701
7042
890
9232
891
8900
4616
991
SM410666
SM410667
SM410668
SM410669
SM410670
SM410671
SM410672
SM410673
SM410674
SM410675
SM410676
SM410677

Что сделано:

- ✓ освоили работу с пакетами R: pathview, gage, GEOquery
- ✓ освоили работу на Phantasus
- ✓ визуализировали интересующие сигнальные пути (реальные данные экспрессии генов в здоровых и патогенных лимфатических эндотелиальных клетках)
- ✓ реализовали алгоритм в R, позволяющий получать данные об экспрессии генов из базы данных GEO (The NCBI Gene Expression Omnibus) и подбирать наиболее значимые сигнальные пути (“pathways”) из KEGG PATHWAY

Каков конечный продукт:

Программа/инструмент, позволяющий пользователю напрямую получать данные об экспрессии генов из базы данных GEO (The NCBI Gene Expression Omnibus) и в автоматическом режиме подбирать наиболее значимые сигнальные пути из KEGG PATHWAY. Пользователю достаточно указать название GSE-файла (GSExxx), выбрать платформу (GPLxxx), на которой были получены данные, указать интересующие образцы (GSMxxxxxx) и выбрать, какие из них считать контрольными, а какие экспериментальными.


Dataset
R,
GEOquery
Количество
значимых
сигнальных
путей
Pathway
database
gage,
pathview
packages/R
pathview image #1
pathview image #2
pathview image #3
etc...