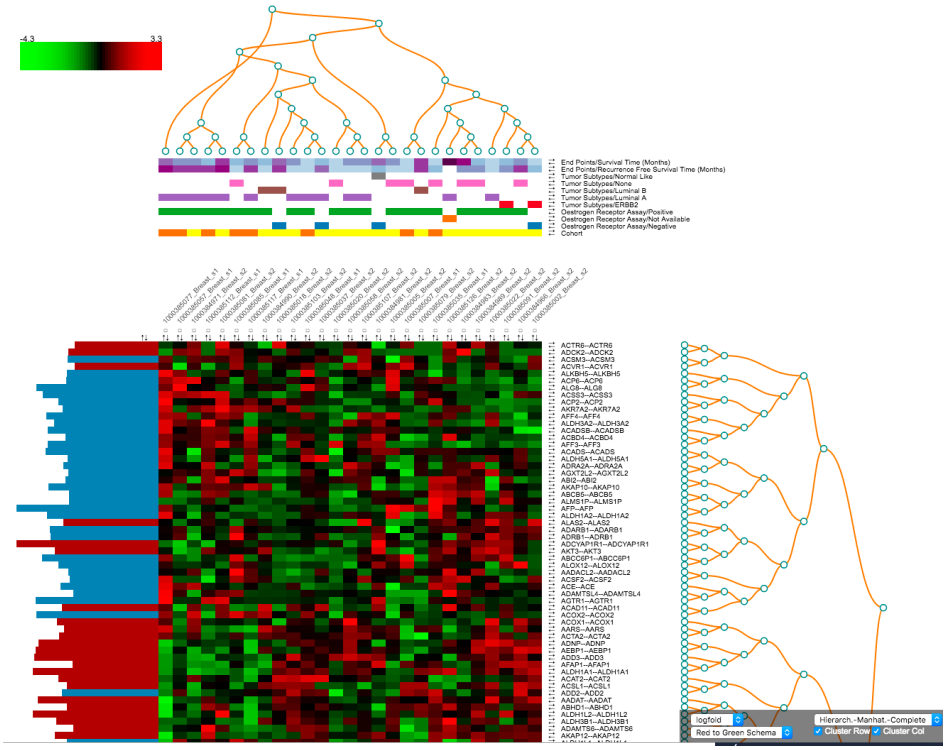
Supplementary Materials

SmartR: An open-source platform for interactive visual analytics for translational research data

Sascha Herzinger, Wei Gu, Venkata Satagopam, Serge Eifes,
Kavita Rege, Adriano Barbosa Da Silva, Reinhard Schneider

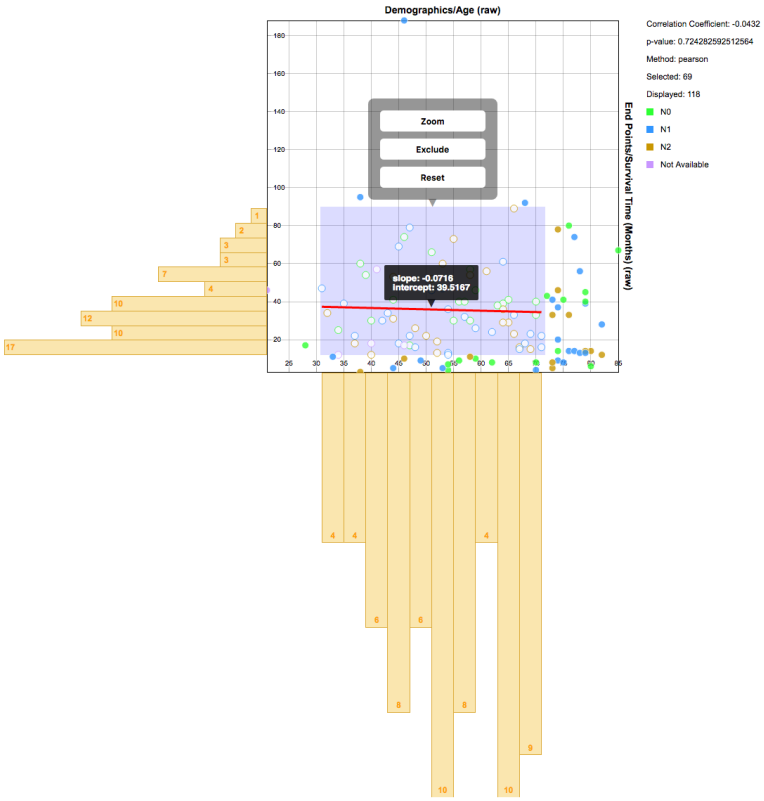
Here we provide a few example implementations of visual analytics using the SmartR framework.
All tools shown in the figures can be accessed via the public server at http://smartR.lcsb.uni.lu

1) Dynamic Heat Map:
For a detailed list of the features for this workflow, please refer to the associated paper. This figure is included for completeness and includes a link to a video that demonstrates the dynamic nature of the heat map.



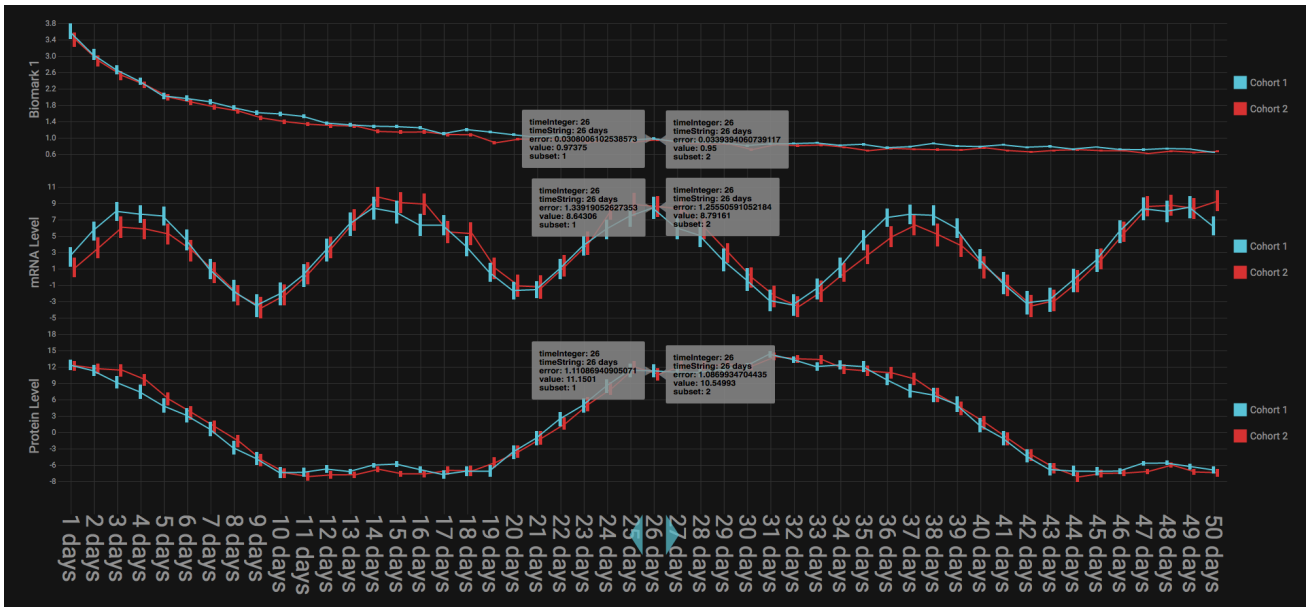
SFig. 1 Dynamic heat map. *The shown heat map is fully sortable and contains many interactive elements that are listed in detail in the associated paper.*
Video URL: *https://youtu.be/kLRSOMbKuns*

2) Correlation Analysis:
This workflow consists of a scatterplot, histograms for the respective axes, and a legend with statistics, such as correlation coefficient and p-value. The scatterplot allows user interaction by area selection, which triggers the re-computation of the correlation statistics, the regression line position, and the histograms. The context menu, available via right-click, reveals the options to zoom, reset or exclude the selected area.



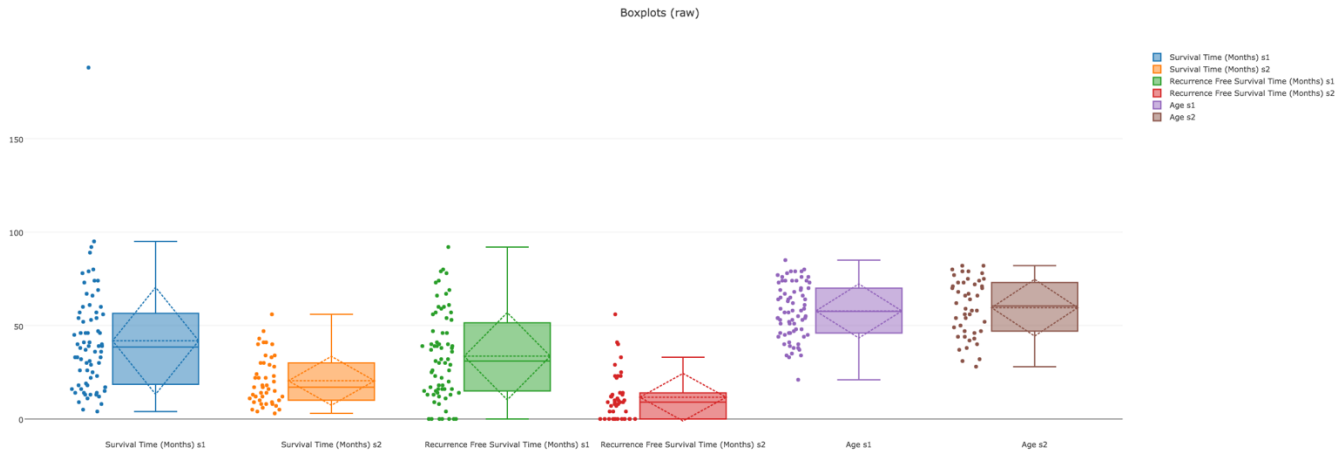
SFig. 2 Correlation Analysis. *Shown is a scatterplot with histograms for the respective axes. Statistics and other plot elements adapt dynamically to certain user-triggered events.*
Video URL: https://youtu.be/IP7vBSQsaFA

3) Line Graph:
The main purpose of the line graph is to visualize longitudinal data. Main features are the manually sortable x-axis, which is important in case the data are not time-annotated (unordered), hover-events that highlight certain data or reveal detailed information (s. SFig. 3), and the possibility to switch instantly between several display modes (e.g. mean with SEM error bars, median with SD error bars, or raw data).



SFig. 3 Line graph. *Shown is a plot that visualizes machine generated time series data. Besides many visual helpers that are triggered by mouse-over events, this visualization has a manual sortable x-axis and different methods for defining the shown timelines.*
Video URL: https://youtu.be/GFCmJysbTyM

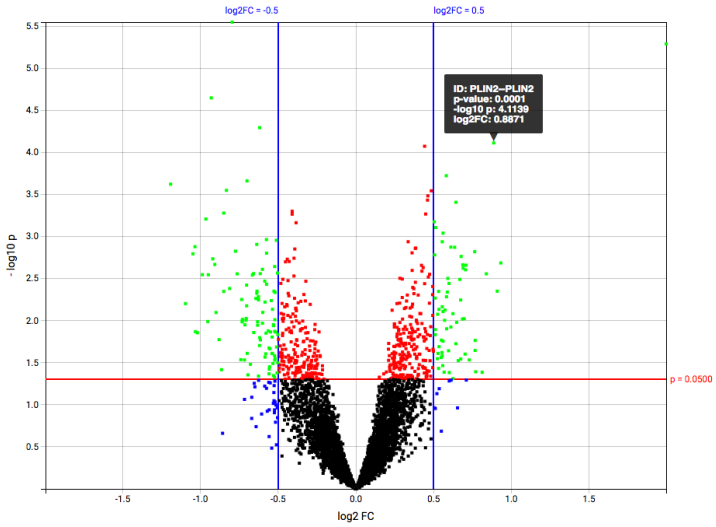
4) Boxplots:
This workflow features boxplots with a diamond-like shape that represents mean with standard deviation. Besides the option to log-transform your data before visualization, it is worth mentioning that the possibility exists to select and treat single rows of micro-array data as normal numerical input for this workflow.



SFig. 4 Boxplots. *Currently this is the only visualization that is using Plotly (Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. https://plot.ly.) as a visualization library. This is included for completeness and to demonstrate that the framework is not limited to certain visualization libraries.*

5) Volcano Plot:

This workflow helps to identify micro-array features that have both a high log-fold change and a significant p-value. Besides a tooltip that shows the statistics for each data point, there is a table that lists all data points with an absolute log-fold change greater than 0.5 and a p-value lower than 0.05. These values are represented by limiters in the scatterplot, which are drag-enabled and instantly update the table on change. Similar to the dendrogram feature in the SmartR heat map, we can launch a KEGG pathway enrichment analysis for genes of interest.



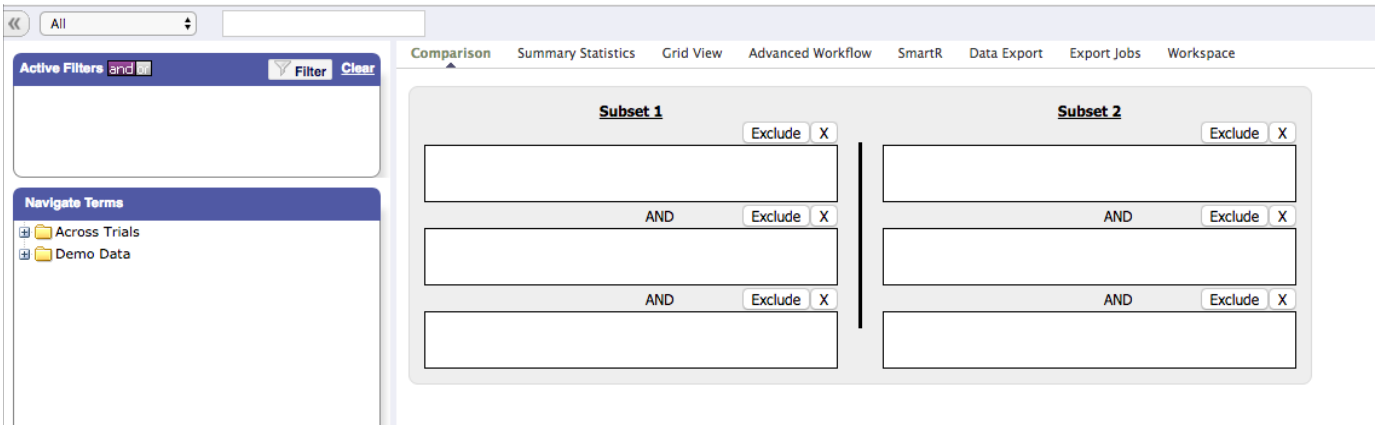
ID	log2 FC	- log10 p	p
CYB5D1--CYB5D1	-0.7963	5.55	0.0000028186
S100A8--S100A8	1.9969	5.2919	0.0000051058
KDM4B--KDM4B	-0.9305	4.6494	0
BHMT--BHMT	-0.619	4.2954	0.0001
PLIN2--PLIN2	0.8871	4.1139	0.0001
SLC4A3--SLC4A3	0.582	3.7237	0.0002
KCNA1--KCNA1	-0.7003	3.6618	0.0002
RERG--RERG	-1.1919	3.6235	0.0002
C2ORF27A--C2ORF27A	-0.8327	3.5498	0.0003
SLC31A1--SLC31A1	0.6441	3.4078	0.0004
STEAP2--STEAP2	-0.85	3.2798	0.0005
BCL2--BCL2	-0.9651	3.2088	0.0006
KIF3C--KIF3C	0.5046	3.176	0.0007
SLC25A5--SLC25A5	0.5143	3.107	0.0008
TTYH3--TTYH3	0.5612	3.041	0.0009
HTT--HTT	-0.5755	2.9643	0.0011
MYO6--MYO6	-0.5113	2.956	0.0011
UBE3C--UBE3C	0.556	2.9388	0.0012
MST1--MST1	-0.6376	2.9075	0.0012
SLC40A1--SLC40A1	-1.0351	2.8799	0.0013
PHACTR2--PHACTR2	0.611	2.8749	0.0013
RBM38--RBM38	0.6356	2.8728	0.0013
MALAT1--MALAT1	-0.7743	2.8264	0.0015
GSTP1--GSTP1	0.765	2.8204	0.0015
C10ORF32--C10ORF32	-0.5748	2.8027	0.0016
GATA3--GATA3	-1.0484	2.7951	0.0016
RNF114--RNF114	0.5071	2.7822	0.0017
GGH--GGH	0.6774	2.7643	0.0017
SPARCL1--SPARCL1	-0.9207	2.7353	0.0018
TRPV6--TRPV6	0.9327	2.6863	0.0021
SYNPO2--SYNPO2	-0.9076	2.6687	0.0021

SFig. 5 Volcano plot. *Shown is the widely known volcano plot that has been enhanced by some dynamic elements. The limiters are drag-enabled and trigger an update for the right-hand table displaying the most significant genes.*

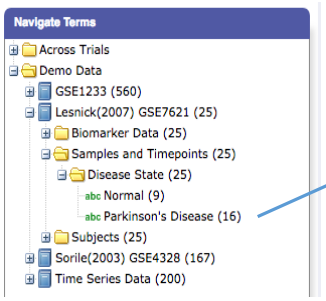
Video URL: https://youtu.be/OxEgDs8vebY

Show Case – Guided Analysis of Parkinson Disease (PD) dataset

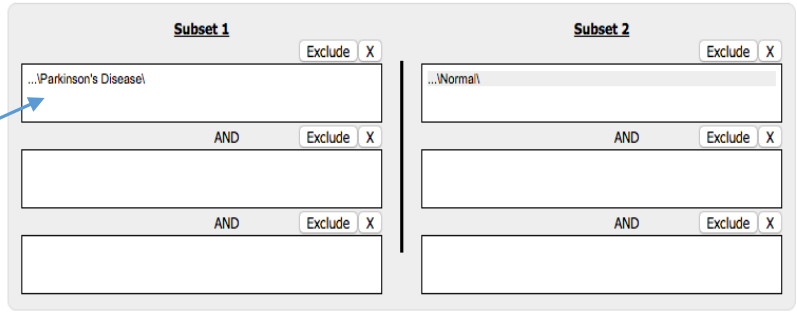
In the following we will demonstrate the capabilities of the SmartR Heat Map in a show-case with the GEO study GSE7621, which contains *"Expression data of substantia nigra from postmortem human brain of Parkinson's disease patients (PD)"* [1]. *The study also contains patients without the disease, serving as a control group.*



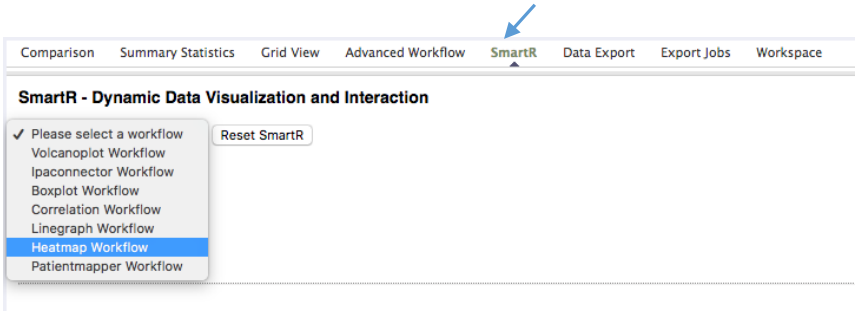
***SFig. 5 The Comparison Tab**. This is our first view when lading the web page. On the left side, our data are represented in a tree-like structure. On the right side we can, based on this data, define case and control group.*



***SFig. 6a The i2b2 Tree.** For this show case, we expand the data tree to expose case and control group of the GSE7621 study as shown in the figure.*

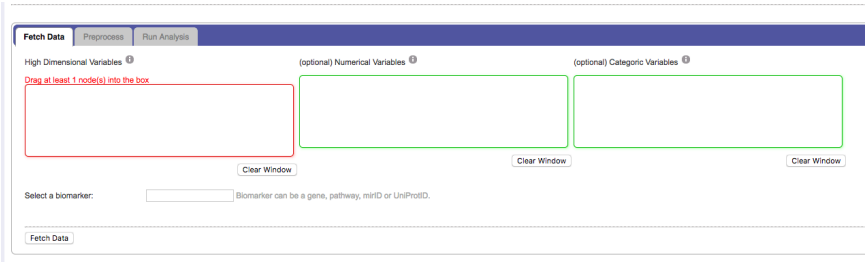


***SFig. 7 Defining case and control group.** By dragging the respective nodes into the two subset windows, we define our cohorts for the following analysis. In our case, we want to compare PD subjects with the healthy control group.*

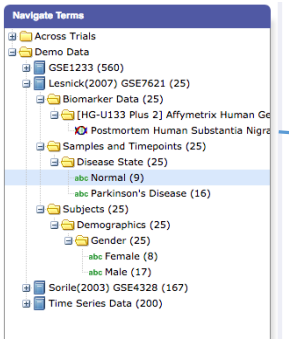


***SFig. 8 The SmartR Tab.** Once our groups are defined we switch to the SmartR tab and select the 'Heatmap Workflow' as shown.*

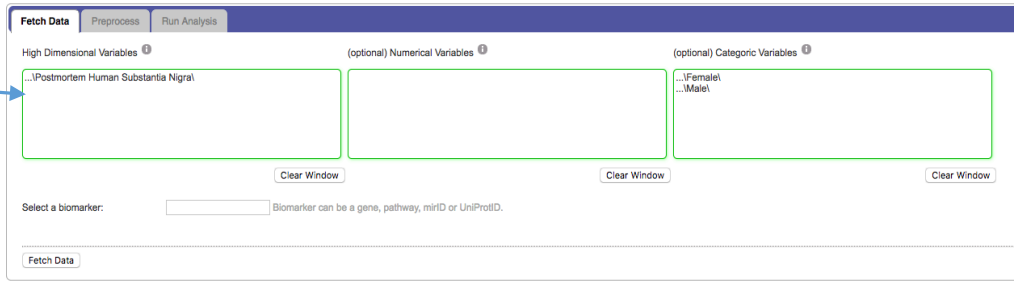
[1] *Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 2007 Jun;3(6):e98. PMID: 17571925*



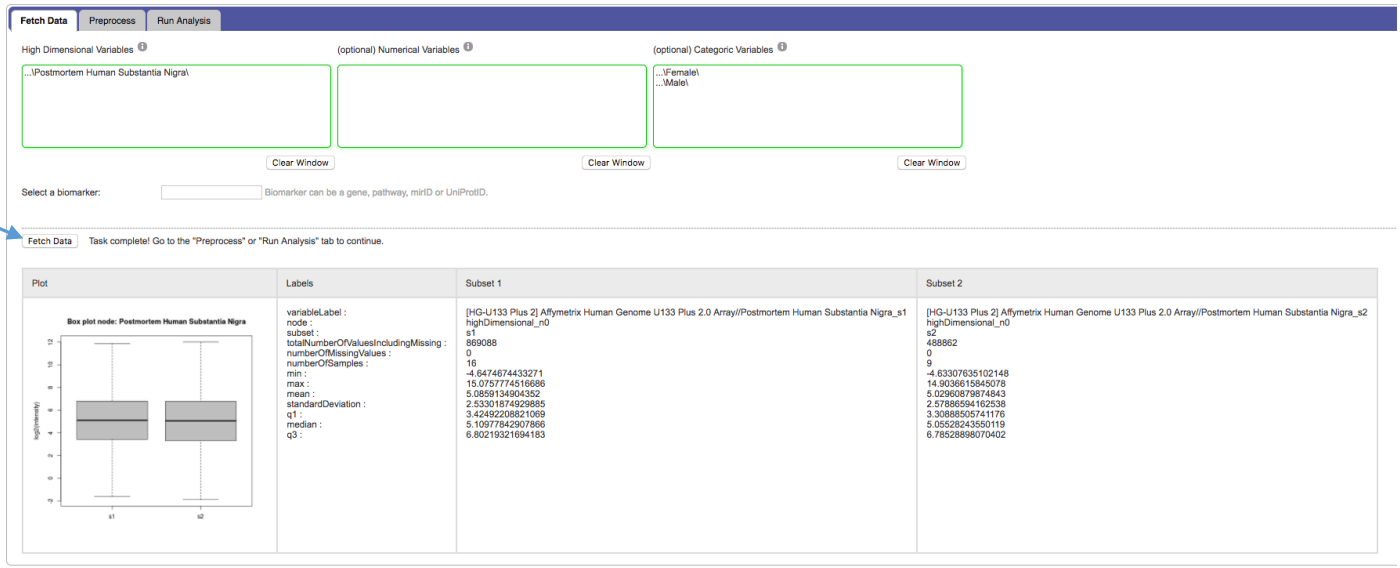
SFig. 9a The Fetch Tab. *After selecting a workflow, we are presented with several boxes. These boxes can be used to assign roles to our data nodes from the tree on the left. In this showcase we obviously must define the expression data we wish to analyze, but are free to include clinical data, as we will see later.*



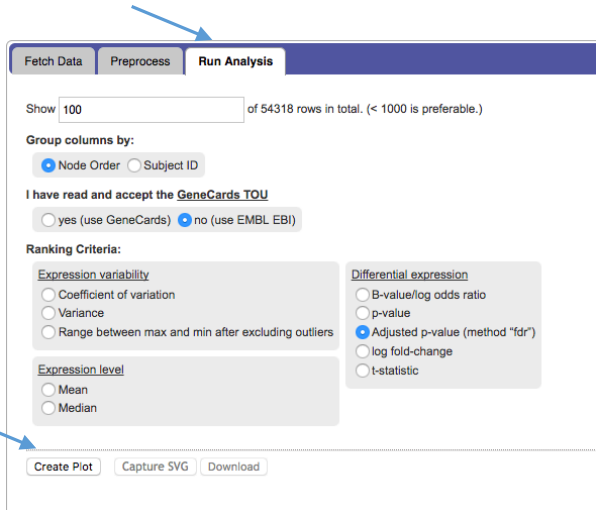
SFig. 6b The i2b2 Tree. *We further expand the tree to show the subject gender and the expression data.*



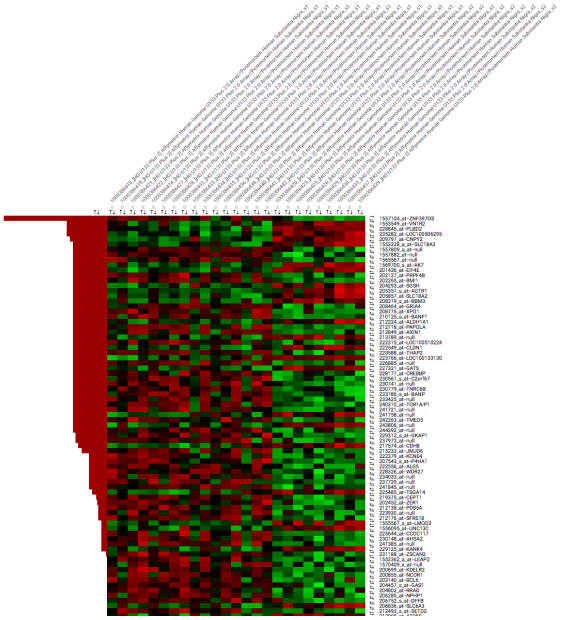
SFig. 9b The Fetch Tab. *Again, by using drag-and-drop, we define what data we wish to include in our analysis. Expression data are dragged to the left-most box and the two categorical gender nodes are dragged to the right-most. As we will see in a moment, this will allow us to relate subjects to their respective gender.*



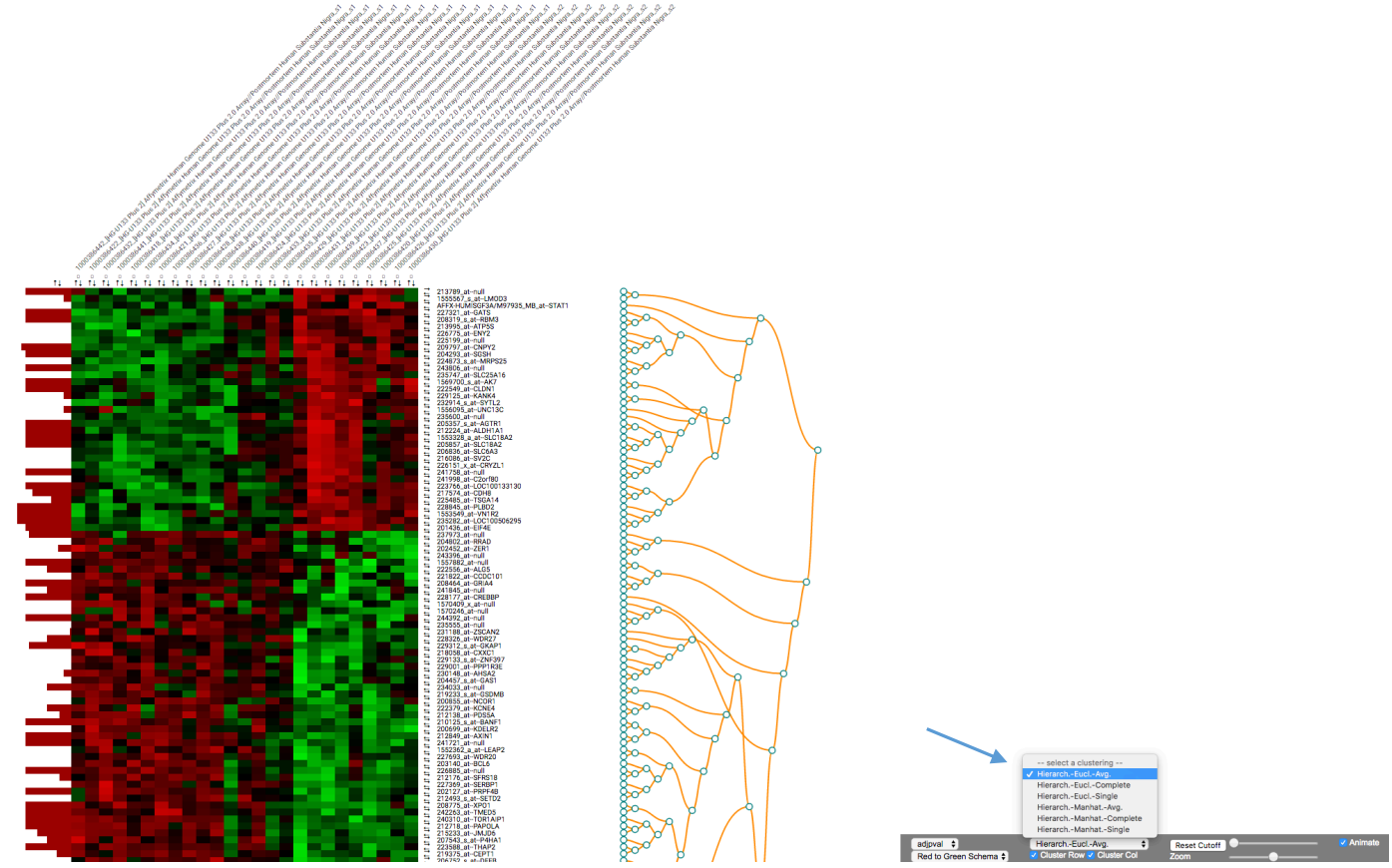
SFig. 9c The Fetch Tab. *Now we press the button "Fetch Data" and wait a moment until all data are loaded and can see a small summary of statistics. In the background this step fetches the just defined nodes from the database and assigns the subjects to their respective cohort that we defined earlier.*



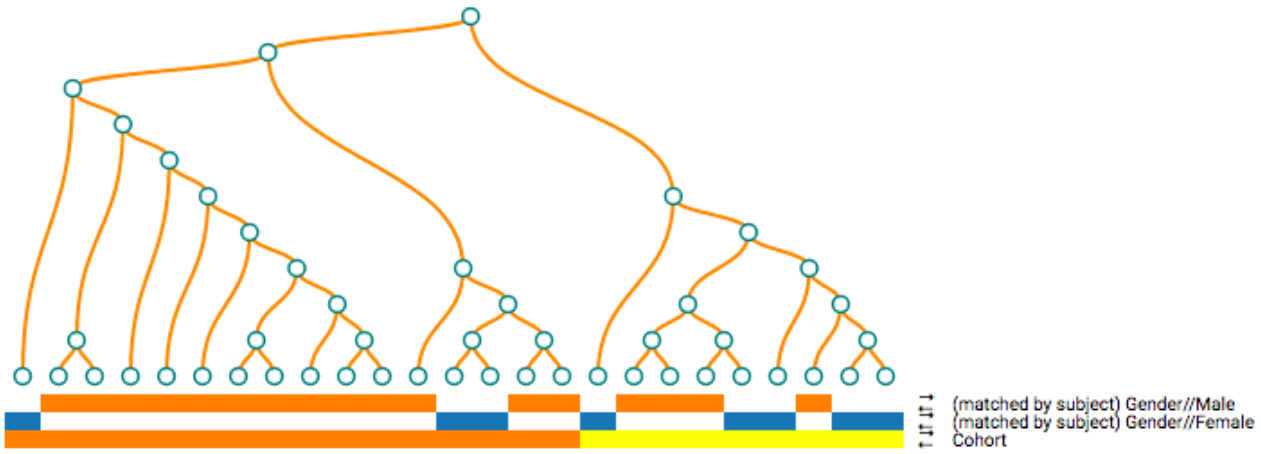
SFig. 10 The Run Analysis Tab. *Now we can switch to the "Run Analysis" tab, where it is possible to set several parameters to tweak the analysis script before creating the heat map. For this show case, the default parameters are a good choice.*



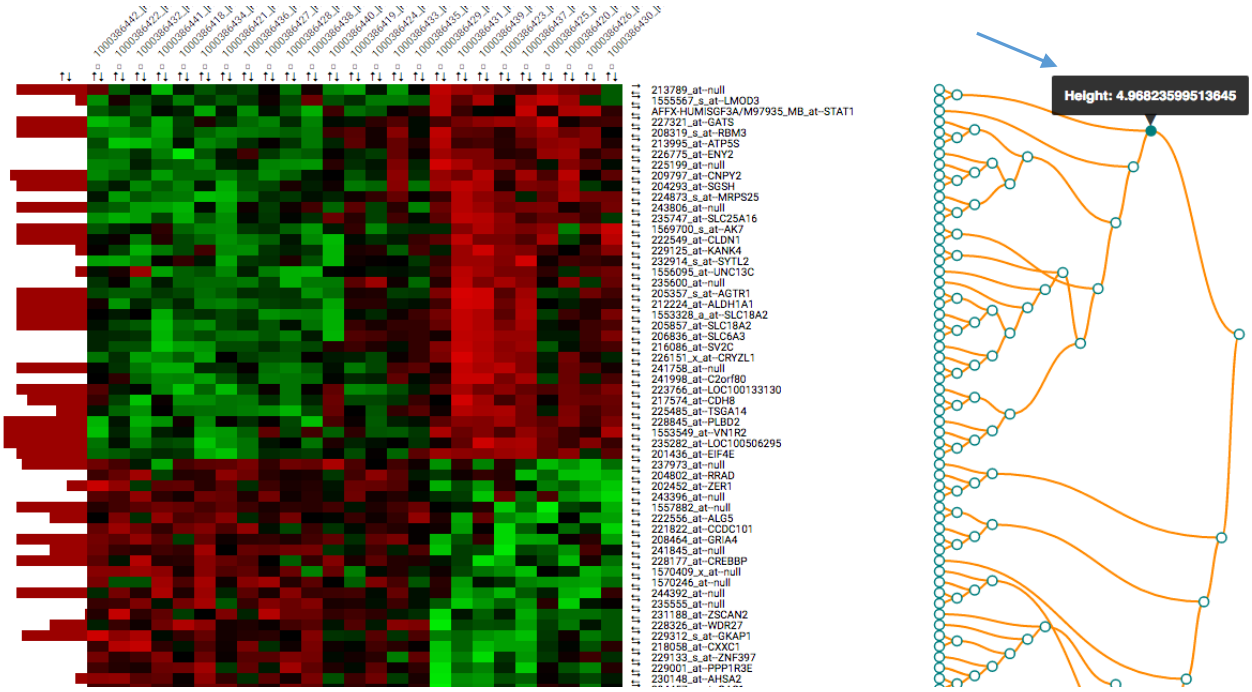
SFig. 11a The Heat Map. *Once the analysis is finished the dynamic heat map is created. Visible are the top genes according to a differential expression analysis executed based on the parameters of the previous figure. Features of the heat map are introduced as needed in the further course of this show case.*



SFig. 11b The Heat Map*. In the lower right corner of the screen we can find several options that will modify our heat map on-the-fly. For now, we want to apply a hierarchical clustering to see if we can clearly separate case and control group. The option "Hierarch.-Eucl.-Avg." will apply such a clustering with 'Euclidean' distance measure and 'Average' linking.*



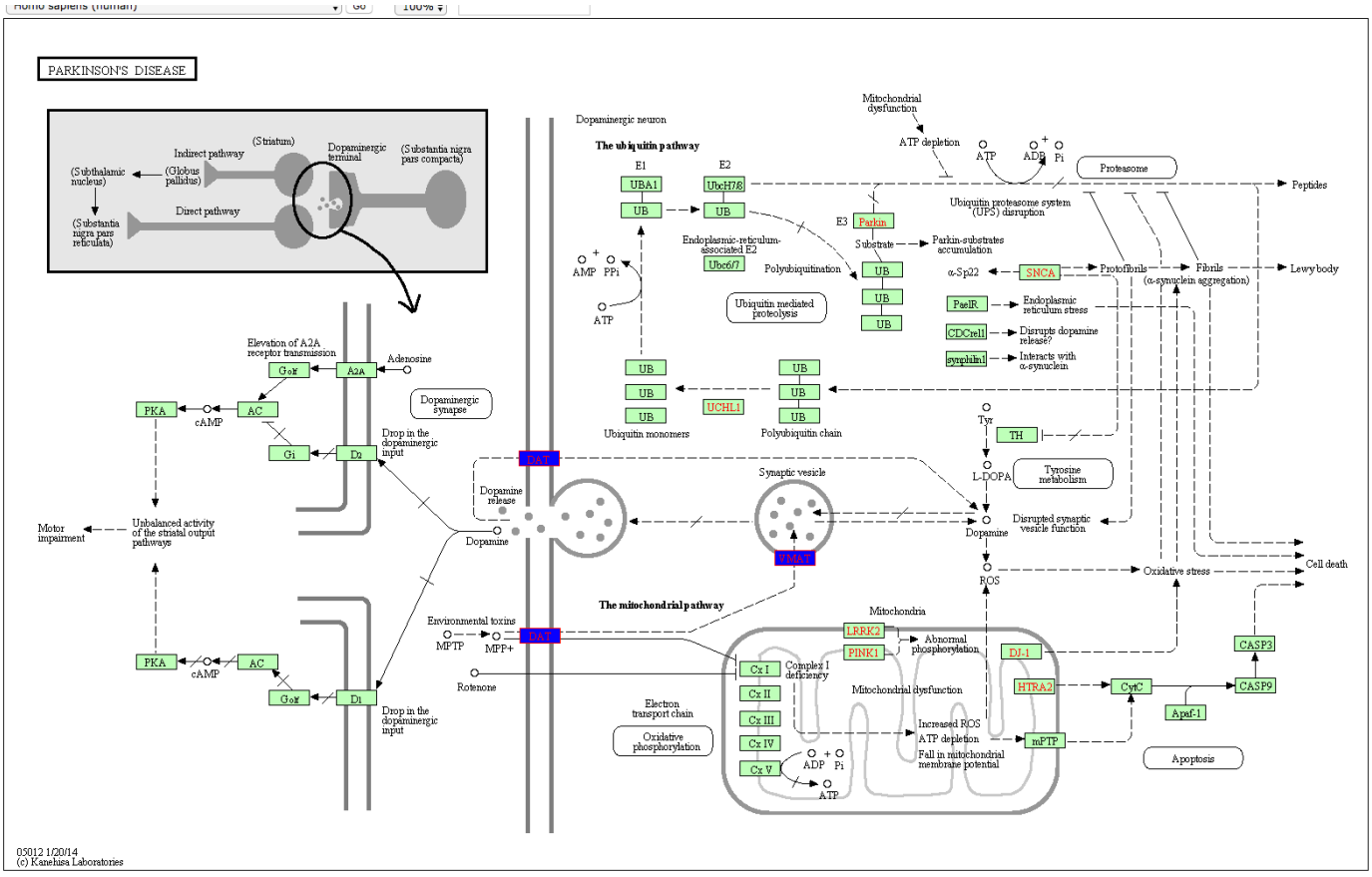
***SFig. 12 The Subject Dendrogram**. Above the heat map we can see our cohorts, encoded in yellow and orange color, and the previously added gender nodes. We can deduce two facts from this image: 1. The clustering could clearly separate our cohorts into two groups. 2. The subject gender seems to have no correlation as to how the data are grouped.*



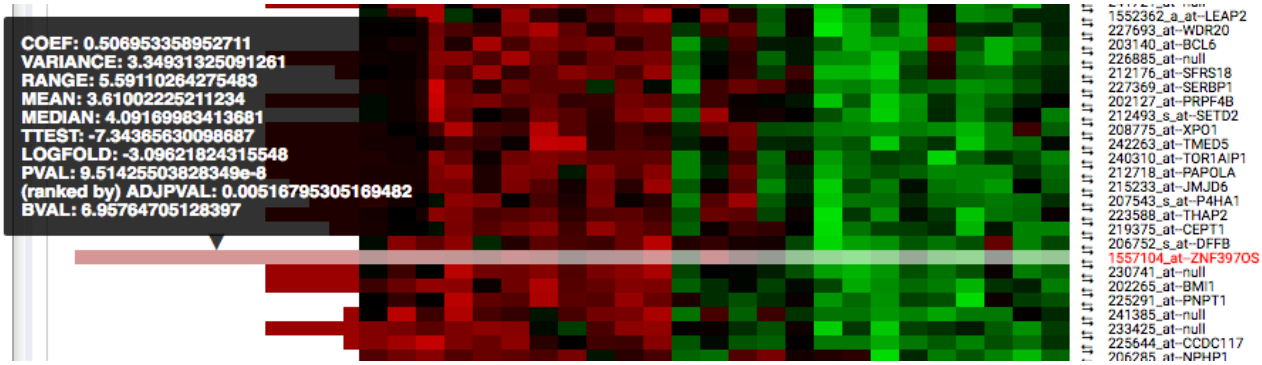
***SFig. 13 The Row Dendrogram**. In a similar fashion the genes are well separated into two sub groups. For further analysis, we might want to attempt to associate these genes with a KEGG pathway. By clicking the marked dendrogram node the contained leafs/genes are send to an external service to handle this task (bioCompendium. The high-throughput experimental data analysis platform (2016). Retrieved from [http://biocompendium.embl.de/)](http://biocompendium.embl.de/).*

KEGG Pathway ID	KEGG Pathway Name	Adjusted P-Value	Gene Name	KEGG Gene	Ensmebl Gene
hsa05012	Parkinson's disease	8.0079e-02	SLC6A3	6531	ENSG00000142319
			SLC18A2	6571	ENSG00000165646
hsa04614	Renin-angiotensin system	1.1301e-01	AGTR1	185	ENSG00000144891
hsa00531	Glycosaminoglycan degradation	1.1301e-01	SGSH	6448	ENSG00000181523
hsa05130	Pathogenic Escherichia coli infection	1.2441e-01	CLDN1	9076	ENSG00000163347
hsa00830	Retinol metabolism	1.2441e-01	ALDH1A1	216	ENSG00000165092
hsa04530	Tight junction	1.2441e-01	CLDN1	9076	ENSG00000163347
hsa04910	Insulin signaling pathway	1.2441e-01	EIF4E	1977	ENSG00000151247
hsa04670	Leukocyte transendothelial migration	1.2441e-01	CLDN1	9076	ENSG00000163347
hsa05212	Pancreatic cancer	1.2441e-01	STAT1	6772	ENSG00000115415

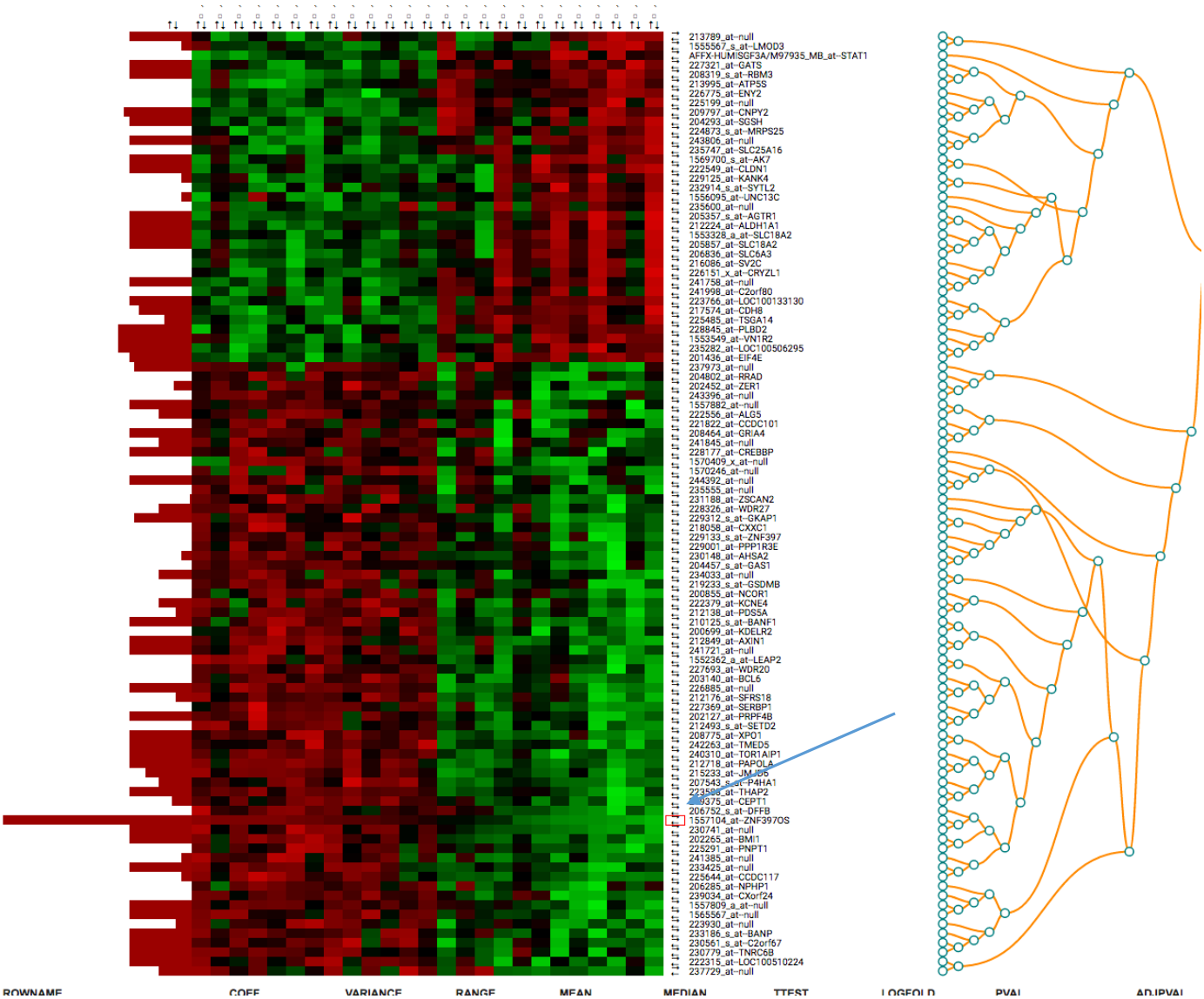
SFig. 14 BioCompendium. *Sure enough, we see an association of these genes with the PD Pathway. Clicking the red flag on the left side will bring us to the KEGG web service.*



SFig. 15 KEGG Pathway. *Associated genes are highlighted in an image of the pathway for further interpretation or analysis.*



SFig. 16 Statistics representation. *Back in the heat map, we have a look at the red bars on the left side. Based on our analysis parameters, these bars initially indicate the adjusted p-value for the respective gene. One gene shows an extraordinary low p-value, represented by a long bar.*



SFig. 17a Manual Sorting. *By clicking the buttons around the heat map, we can manually sort by columns or rows. Sorting the columns based on the expression values for this single gene seems to have almost no impact.*

SFig. 17b Manual Sorting. *Sure enough, the previous observation is confirmed by the nearly perfect split between case and control group. This makes the gene a good candidate for further analysis. If one wishes to proceed from here, clicking the name of the gene will redirect the analyst to another service, where further gene specific details are listed.*

ROWNAME	COEF	VARIANCE	RANGE	MEAN	MEDIAN	TTEST	LOGFOLD	PVAL	ADJPVAL	BVAL
213789_at--null	0.352943200821252	1.7032880290322	4.83289001416474	3.69776431584249	3.62527048937469	4.97593734827021	1.89137878605924	0.00003760503336889	0.0869878807677252	2.08530185818681
1555567_s_at--LMOD3	0.818551884240798	1.793393949269194	4.30597052098439	1.63603120226232	1.11769504266975	4.48241719411219	1.83450543238813	0.000136816473212012	0.120828189773694	0.994827641090466
AFFX-HUMISGF3A/M97935_MB_at--STAT1	0.0781214728281592	0.197309996498668	1.52040197014876	5.68596421323868	5.68486544302731	4.31852418711318	0.646174899053455	0.000209947585326376	0.127647417596733	0.631615795062202
227321_at--GATS	0.0908692008599782	0.459546185268008	2.65241829450611	7.46015539569636	7.40122086216022	5.12021470848492	1.02705232101601	0.000025796953283767 8	0.0869878807677252	2.4017345087356
208319_s_at--RBM3	0.252587739322224	2.89833574726434	5.95547882481457	6.74003389715966	6.80748372836463	4.93483976721954	2.44517119931819	0.000041871408865 7852	0.0869878807677252	1.99490897436962
213995_at--ATP5S	0.110033512301576	0.673601915557448	3.12929830045997	7.45893256482656	7.61102479730735	4.41545691645788	1.13484832952241	0.000162992225457354	0.127647417596733	0.846419674611436
226775_at--ENY2	0.0735141401893476	0.24670965579727	1.25868838005865	6.75650638525501	6.78541977214973	4.28988762048966	0.70752112559484	0.000226234530613582	0.127647417596733	0.56818097637126
225199_at--null	0.0511266740331572	0.213193443614 88	1.73662801306176	9.03107358707673	8.94941820539521	4.37754668755168	0.672674139427027	0.000179963410526989	0.127647417596733	0.762399969024467
209797_at--CNPY2	0.0708552178764762	0.286432587016272	1.74439103315112	7.55334592756206	7.4297833205085	5.58543816430386	0.864988403094491	0.000007702446208860 98	0.0836762946345822	3.40977324927344
204293_at--SGSH	0.51652910517626	1.7553591323153	4.60627300696657	2.56500486709472	2.29865831556452	4.86443882817427	1.89722530305374	0.000050338857029024	0.0869878807677252	1.83983336255689
224873_s_at--MRPS25	0.0663973188002247	0.169165719982571	1.49951746918396	6.19449117160491	6.14526985665051	4.30182118446893	0.605289709879719	0.00021930019195792	0.127647417596733	0.594613943741632
243806_at--null	0.418988830263621	2.63355405193145	5.68182403997375	3.87318884900711	4.12018602753186	5.46953852975785	2.45196492438806	0.000010396672335947 5	0.0869878807677252	3.16067211180606
235747_at--SLC25A16	0.0805445819778756	0.324017867177399	2.0332499801206	7.06721245987517	7.02691069122452	4.28381046308573	0.796566209495246	0.000229849066706 48	0.127647417596733	0.554721140302714
1569700_s_at--AK7	0.392845195294658	1.63581186305503	4.37889033307107	3.25570638192988	2.85798099512757	4.89511338337228	1.83894181925115	0.000046456614516503 3	0.0869878807677252	1.90743585092802
222549_at--CLDN1	0.109542355914723	0.315422915669389	2.11214727932095	5.12701447923913	5.08406426478848	4.79316452269337	0.837656380467608	0.000060661882280774 3	0.0869878807677252	1.68256941523489
229125_at--KANK4	0.0920480024396105	0.514134641732407	2.8570340131693	7.78975978353561	7.78463484555752	4.46051963271716	1.005787188441	0.000144879672280656	0.121070369829856	0.946295953450335
232914_s_at--SYTL2	0.0554568295567218	0.192750772682419	1.39073094988046	7.91667882677067	7.95070169008901	4.28783980214826	0.637298850071838	0.000227446170772471	0.127647417596733	0.563645330408717
1556095_at--UNC13C	0.303535638653324	3.58184067031265	6.89611784292279	6.23510027019202	5.99118168140853	4.47034343363058	2.57454080842771	0.000141205374308935	0.120828189773694	0.968068784974647
235600_at--null	0.376450161060039	2.57838844077629	5.86904884213189	4.26546797340319	3.91934008244201	4.39419013902884	2.16788389744967	0.000172306359133253	0.127647417596733	0.799285540286965
205357_s_at--AGTR1	0.275898853402348	2.13004431266795	5.10173309332926	5.28986298828583	4.64270157183323	5.3142587546585	2.1795005520628	0.000015559425408088	0.0869878807677252	2.82471751289301
212224_at--ALDH1A1	0.155575975008584	1.94422010511397	4.44735146802301	8.96252094273373	8.65971053515448	4.77525431528467	1.97529345266641	0.000063573607557932 2	0.0869878807677252	1.64301403086427
1553328_a_at--SLC18A2	0.399271807119352	4.52525517314364	8.002222397637	5.32786106695321	5.13791332208841	5.27447823066081	3.14992569426341	0.000017256202809458	0.0869878807677252	2.73827264697508
205857_at--SLC18A2	0.307302071798484	4.5227768370151	8.20136303727345	6.92049391975417	6.66391384211598	4.96042304027279	3.05505688903249	0.000039161776139675 5	0.0869878807677252	2.05119093529373
206836_at--SLC6A3	0.248468403309125	3.42764586256496	6.79676811383928	7.45120999946554	7.05539090473479	4.40105291544567	2.49701932599188	0.000169244229574882	0.127647417596733	0.814495476126659
216086_at--SV2C	0.176874287141326	2.363452248287 17	6.2518962279749	8.6917797658661	8.50314993743244	4.38652585073807	2.07493663515307	0.000175791308563416	0.127647417596733	0.782299542613199
226151_x_at--CRYZL1	0.0459749656696201	0.153187272667031	1.27430414245455	8.51314299791171	8.41438926800622	4.28221013563674	0.580169401921816	0.000230810399700418	0.127647417596733	0.551176829011476
241758_at--null	0.180459793928244	1.20428899598069	4.0070044637626	6.08113860087157	5.9518675043269	4.89204716924447	1.58407437565897	0.000046830781256708 6	0.0869878807677252	1.90068064578951
241998_at--C2orf80	0.157043496308869	1.03188055545553	3.70301744178367	6.46836857976374	6.23515262421793	4.39578013311565	1.38733541113044	0.000171592043462583	0.127647417596733	0.802809413318931
223766_at--LOC100133130	0.0900260084207171	0.329201517611477	2.03861195216509	6.37327883453007	6.4178525148859	5.21557405503171	0.890947229545502	0.000020117270808796 6	0.0869878807677252	2.61000772946124
217574_at--CDH8	0.144159422846422	0.819089681618605	3.95937230803183	6.27801996654448	6.30578816681099	4.69933954201581	1.28870311841505	0.000077551030767194 8	0.093609264204722	1.47520949020573
225485_at--TSGA14	0.0881896009349739	0.356908710660954	2.06873449596941	6.77424963048349	6.76261457875456	4.56803583995353	0.862570857605389	0.000109359702724317	0.110003709862582	1.18454246557575
228845_at--PLBD2	0.158443475958295	0.776837979799648	2.96249710198486	5.56276692326424	5.62175885670973	5.82531771735895	1.39382654832684	0.000004153061523821 07	0.0800919778355269	3.92036137731063
1553549_at--VN1R2	0.514297363285467	1.90935538724509	4.91196807499758	2.68676132207083	2.59693514238723	5.68888584844225	2.131126314306	0.000005898006394604 14	0.0800919778355269	3.63081991197464
235282_at--LOC100506295	0.100990458796967	0.414189254166524	2.34622100818931	6.37263535155339	6.3017705324551	5.78111297351377	1.0359361112711	0.000004621632556819	0.0800919778355269	3.82680348344
201436_at--EIF4E	0.0484584033176724	0.167848450706 38	1.46028631033868	8.45451644600656	8.31745783860680	5.16908707346729	0.666640173073152	0.000022656674888664	0.0869878807677252	2.51053601037780

SFig. 18 Statistics Table. *Finally, most of the data are displayed below the heat map in table form. The table automatically adapts to the current sorting of the heat map rows to increase visibility. These data are also retrievable by clicking one of the buttons near the "Create Plot" button.*