

# eTRIKS : Tutorial for gene expression data analysis

---

Follow up details:

Name	Organization	Date	Details
N. Jullian	CNRS	25/11/2013	version 1.0
N. Jullian	CNRS	05/11/2014	version 2.0

## Introduction

This tutorial focuses on gene expression data analysis using the eTRIKS training server. The objective of the tutorial is to familiarize you with some basic tools available in eTRIKS. The tutorial is based on tranSMART v1.2 (released September 2014) accessible via the following URL: [public.transmart.etriks.org](http://public.transmart.etriks.org)

The dataset used for this example is from GSE10500. The study reports gene expression profiling of synovial macrophage for 5 Rheumatoid Arthritis patients and 3 control subjects. We are going to run scripts for selecting genes differentially expressed in RA patients as compared to control patients.

Launch eTRIKS by typing the following URL <http://public.transmart.etriks.org> in your favorite Internet browser. Go to the Dataset Explorer Tab and open the tree under **Public Studies**. Public datasets that are uploaded into eTRIKS are labeled with pathology name, author name and GSE code.

Expand the nodes by clicking on the + sign left of **GEO Studies** and then **Rheumatoid\_Arthritis** node. The Yirilina study refers to GSE10500. For more details on the study, right click on the study name and select « **Show Definition** ». A popup window shows the summary of the study as deposited in the NCBI repository (Figure 1).

Show Concept Definition-Yarilina(2008)_GSE10500 (8)		AND	Exclude	X
<b>TITLE:</b>	Gene expression in rheumatoid arthritis synovial macrophages			
<b>SUMMARY:</b>	Macrophages from RA synovial fluids were compared to primary human blood-derived macrophages.			
<b>CONTRIBUTOR:</b>	Lionel,B,Ivashkiv; Taras,T,Antoniv; Xiaoyu Hu; Yarilina Anna; Kyung-Hyun Park-Min			
<b>CONTACT:</b>	Name: Taras Antoniv; Laboratory: Lionel B. Ivashkiv, MD; Department: Research; Institute: Hospital for Special Surgery; Address: 535 East 70th Street; City: New York; State: NY; Zip/postal_code: 10021; Country: USA			
<b>TYPE:</b>	Expression profiling by array			
<b>OVERALL_DESIGN:</b>	<p>Macrophages from synovial fluids of five RA patients were isolated by ficoll density gradient centrifugation, followed by positive selection of CD14+ cells using magnetic beads. The patients' diagnoses were determined by their physicians, in each case a Board-certified rheumatologist at the Hospital for Special Surgery, and were definite RA according to American College of Rheumatology criteria. De-identified samples were processed in an anonymous manner using a protocol approved by the Institutional Review Board of the Hospital for Special Surgery.; Human peripheral blood mononuclear cells were isolated from venous blood of independent healthy donors by centrifugation on a Ficoll density gradient. Monocytes were obtained from PBMCs using anti-CD14 magnetic beads, and were &gt;96% pure as verified by flow cytometry. Monocytes were differentiated into macrophages by culturing in RPMI 1640 medium supplemented with 10% fetal bovine serum and M-CSF (20 ng/ml). The viability and purity of macrophages was comparable in all conditions; Gene expression was analyzed using Affymetrix microarrays and protocols.</p>			
<b>STATUS:</b>	Public on Feb 14 2008			
<b>SUBMISSION_DATE:</b>	2008-02-12			
<b>LAST_UPDATE_DATE:</b>	2012-06-08			
<b>PUBMED_ID:</b>	18345002			
<b>ORGANISM(S):</b>	Homo sapiens			
<b>PLATFORM:</b>	GPL8300			
				Close

Figure 1: Study Definition view

## Lesson 1 : Marker selection

### Navigating the data tree

Expand the **Yarilina(2008)\_GSE10500** node by clicking on the + sign on the left. The right number in parenthesis represents the number of subjects in the study, that is 8 here. Now you have 2 nodes listed :

- **Biomarker Data (8)**
- **Samples and Timepoints (8)**

Each node is related to a type of data and we are going to browse through each node to see what kind of data is available for this study (Figure 2).

A label is assigned to each data node :

« **abc** » for categorical text variables

«  » for high dimensional data.

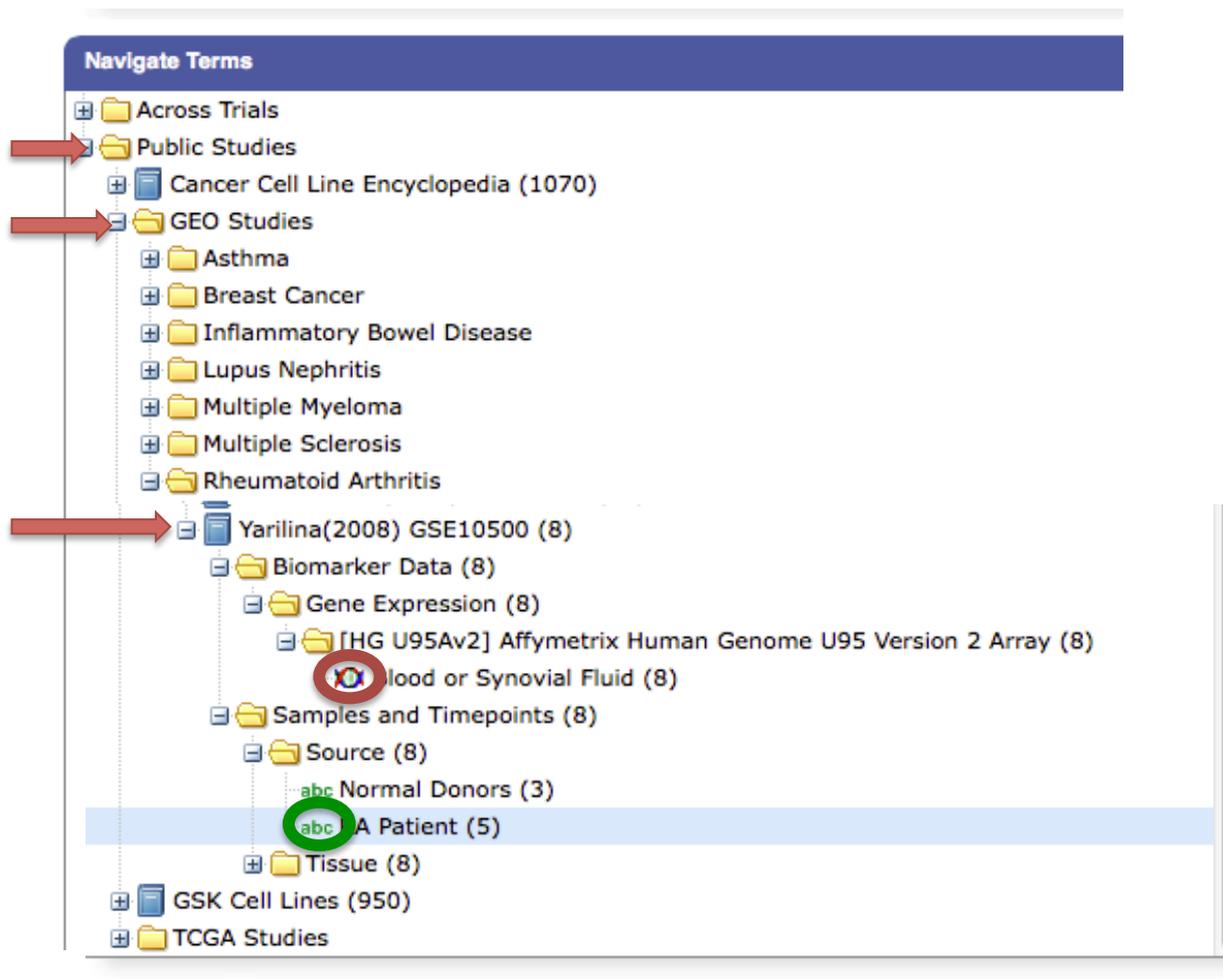


Figure 2: GSE10500 Data Tree View

First, expand the **Biomarker Data** node to access the details of the micro-array based profiling: **HG-U95Av2**, Affymetrix Human Genome U95 version2 Array. Again the number in parenthesis indicates that the experiment has been performed for a total of 8 subjects.

Then, expand the **Samples and Timepoints** node. There are 2 groups of data collected in this study. The **Source** node is related to the clinical characterization of the subjects as **Normal Donor** subjects and **RA Patient** (Figure 2).

**Clinical question** : Can we identify markers able to discriminate between “RA patients” and “Normal Donors” ?

### Define 2 subsets

We are going to start by defining the 2 groups that we wish to compare. Drag the **Normal Donors** node into the « Subset 1 » box and the **RA patient** node into the « Subset 2 » box as shown in Figure 3 by the red arrows.

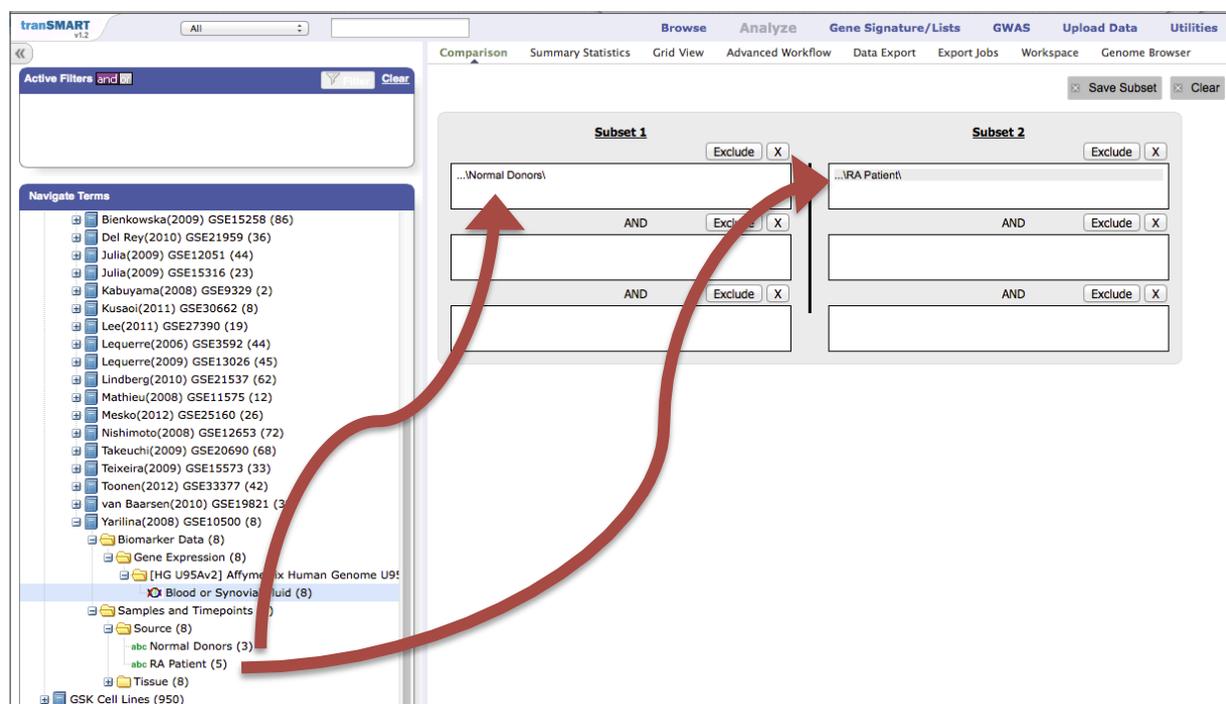
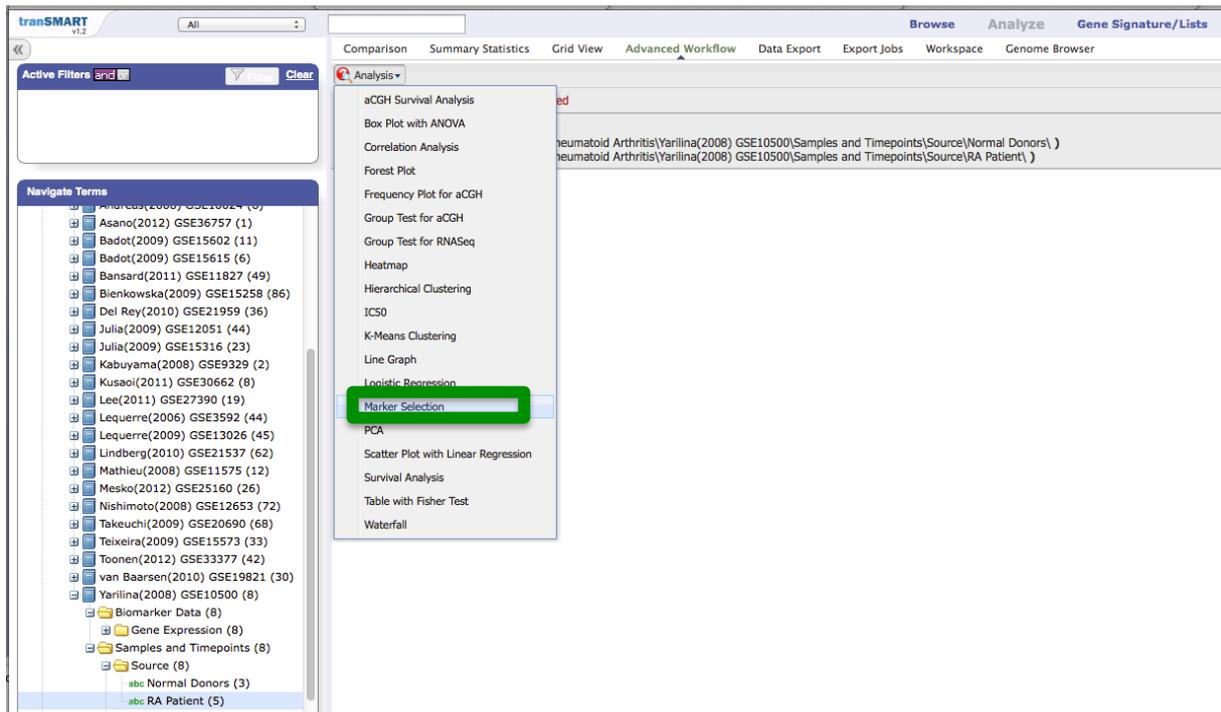


Figure 3: Subset definition window

### Select Analysis workflow

Then, click on **Advanced Workflow** and select « **Marker Selection** » in the **Analysis** pull-down menu (Figure 4a). Drag the High dimensional data node « **Blood or Synovial Fluid** » into the Marker variable box (red arrow, Figure 4b) and click on « **High Dimensional data** » at the bottom right of the box (green rectangle, Figure 4b).

a)



b)

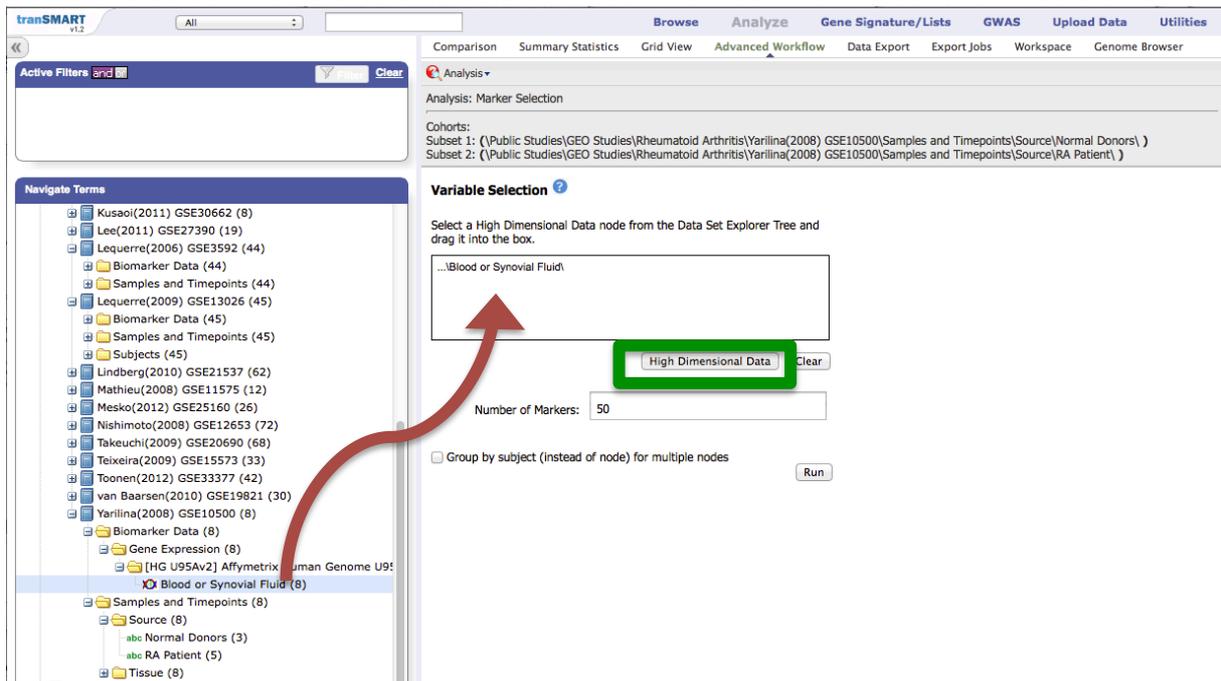


Figure 4: a) Analysis pulldown b) Variable selection box

A popup window lists the characteristics of the Gene Expression platform (Figure 5). Click on **Apply Selections** to validate the default parameters.

Figure 5: default parameters for subset selection

Then, enter « 20 » for the number of markers (Figure 6) :

Figure 6: Marker selection options

Then click on the Run box to launch the calculations. When it is finished, the heatmap (Figure 7) and the table with the list of 20 top markers (Figure 8) are appended at the bottom of the window.

### Visualize the results

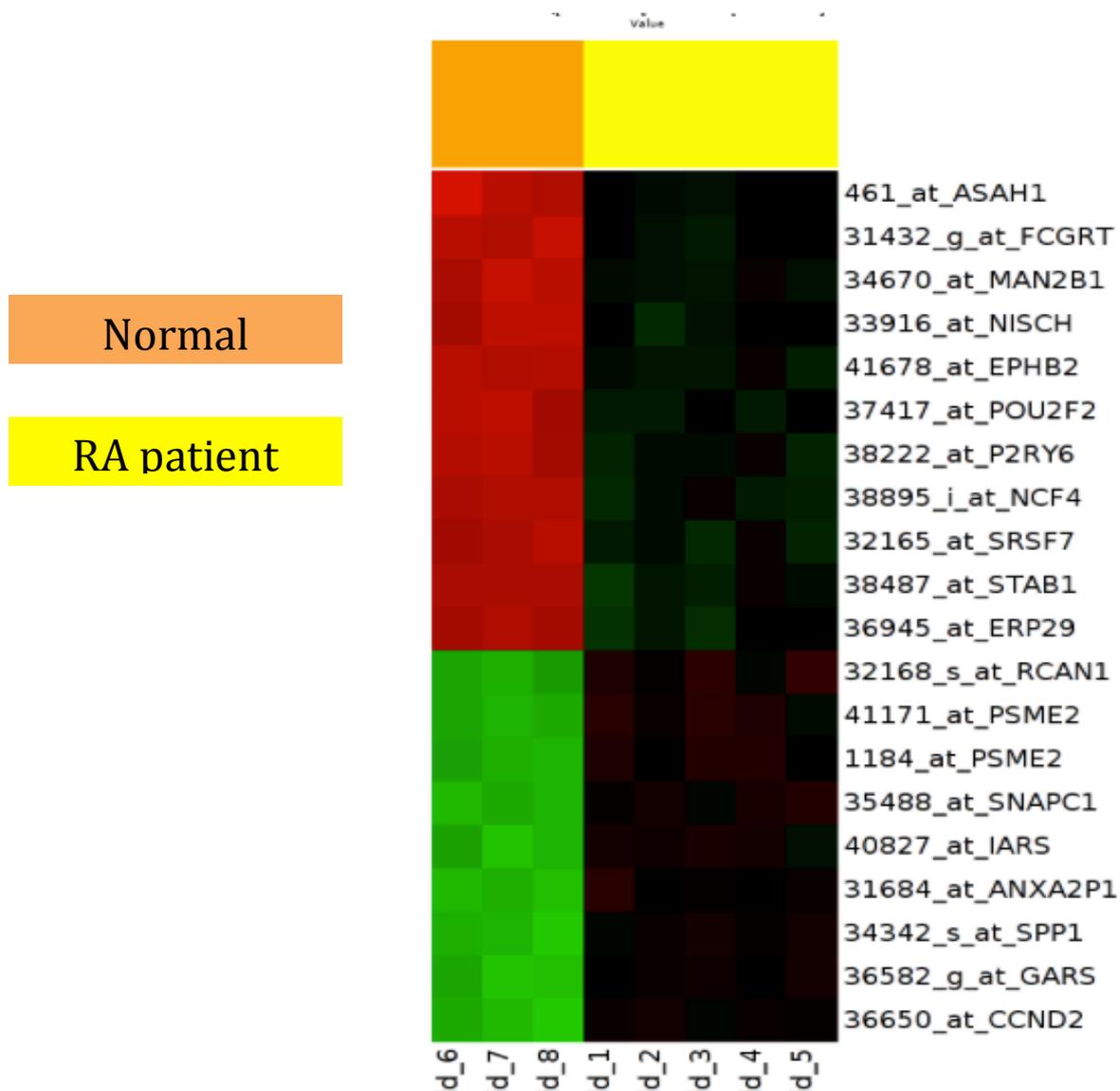


Figure 7: heatmap with 20 most discriminant markers

The heatmap is based on the most discriminant markers between Subset 1 (Normal Donors) and subset 2 (RA patients). The table lists details for the top 20 markers sorted by z-score.

Table of top Markers

Gene Symbol	Probe ID	Raw p-value	Bonferroni	Holm	Hochberg	Sidak	Sidak	Benjamini	Benjamini	t	t (permutation)	Raw P (permutation)	Adjusted P (permutation)	Rank	S1 Mean	S2 Mean	S1 SD	S2 SD	Fold Change (relative to S2)
ANXA2P1	31684_at	0	0	0	0	0	0	0	0	22.95447	22.95447	0.01785714	0.01785714	3	-1.799723	0.11608420	0.06733974	0.16513967	3.773249
ASAH1	461_at	0	0	0	0	0	0	0	0	-17.02578	-17.02578	0.01785714	0.07142857	14	1.858677	-0.06084625	0.18676848	0.07359715	-3.782981
CCND2	36650_at	0	0	0	0	0	0	0	0	20.81155	20.81155	0.01785714	0.03571429	6	-1.829664	0.09080207	0.13548382	0.10947061	3.785452
EPHB2	41678_at	0	0	0	0	0	0	0	0	-24.50289	-24.50289	0.01785714	0.01785714	1	1.770279	-0.14676700	0.05312194	0.16094199	-3.776490
ERP29	36945_at	0	0	0	0	0	0	0	0	-16.71266	-16.71266	0.01785714	0.07142857	16	1.655023	-0.24257505	0.04508329	0.24712758	-3.725923
FCGRT	31432_g_at	0	0	0	0	0	0	0	0	-22.51010	-22.51010	0.01785714	0.01785714	4	1.820855	-0.09934607	0.11242190	0.12377138	-3.784757
GARS	36562_g_at	0	0	0	0	0	0	0	0	19.90074	19.90074	0.01785714	0.03571429	8	-1.824204	0.09643532	0.14848727	0.09911807	3.785909
IARS	40827_at	0	0	0	0	0	0	0	0	15.35179	15.35179	0.01785714	0.10714286	20	-1.775863	0.13266866	0.17245003	0.16646835	3.754268
MAN2B1	34670_at	0	0	0	0	0	0	0	0	-18.27536	-18.27536	0.01785714	0.05357143	11	1.813329	-0.10384171	0.15474185	0.12294941	-3.776816
NCF4	38895_l_at	0	0	0	0	0	0	0	0	-20.06591	-20.06591	0.01785714	0.03571429	7	1.701496	-0.20649212	0.03763468	0.20699314	-3.752853
NISCH	33916_at	0	0	0	0	0	0	0	0	-16.11580	-16.11580	0.01785714	0.08928571	17	1.778232	-0.12932613	0.14850762	0.18246844	-3.751737
P2RY6	38222_at	0	0	0	0	0	0	0	0	-16.87183	-16.87183	0.01785714	0.07142857	15	1.729183	-0.17585475	0.11378155	0.20536704	-3.745446
POU2F2	37417_at	0	0	0	0	0	0	0	0	-17.46489	-17.46489	0.01785714	0.07142857	12	1.750466	-0.16292872	0.14969794	0.15054619	-3.766945
PSME2	1184_at	0	0	0	0	0	0	0	0	19.66325	19.66325	0.01785714	0.03571429	9	-1.715107	0.19651064	0.09351257	0.18078213	3.762308
PSME2	41171_at	0	0	0	0	0	0	0	0	18.79551	18.79551	0.01785714	0.03571429	10	-1.893033	0.21348379	0.06715236	0.20959314	3.749029
RCAN1	32168_s_at	0	0	0	0	0	0	0	0	15.44045	15.44045	0.01785714	0.10714286	19	-1.646249	0.24984923	0.09552059	0.24534277	3.722052
SNAPC1	35486_at	0	0	0	0	0	0	0	0	23.16531	23.16531	0.01785714	0.01785714	2	-1.752604	0.16389567	0.07127522	0.16048526	3.775059
SPP1	34342_s_at	0	0	0	0	0	0	0	0	21.37555	21.37555	0.01785714	0.01785714	5	-1.822786	0.09798289	0.13013256	0.11021960	3.786248
SRSF7	32165_at	0	0	0	0	0	0	0	0	-15.84566	-15.84566	0.01785714	0.08928571	18	1.690497	-0.21161369	0.12420465	0.21525914	-3.737597
STAB1	38487_at	0	0	0	0	0	0	0	0	-17.16562	-17.16562	0.01785714	0.07142857	13	1.675218	-0.22312510	0.01822927	0.24616393	-3.727848

Figure 8: top 20 markers selected

**Conclusion:** This tutorial shows how to define 2 subsets of patients and to extract differentially expressed genes from a high dimensional data study. The report lists the top 20 genes that discriminate the most between the 2 groups, namely « RA patients » and « healthy donors ».