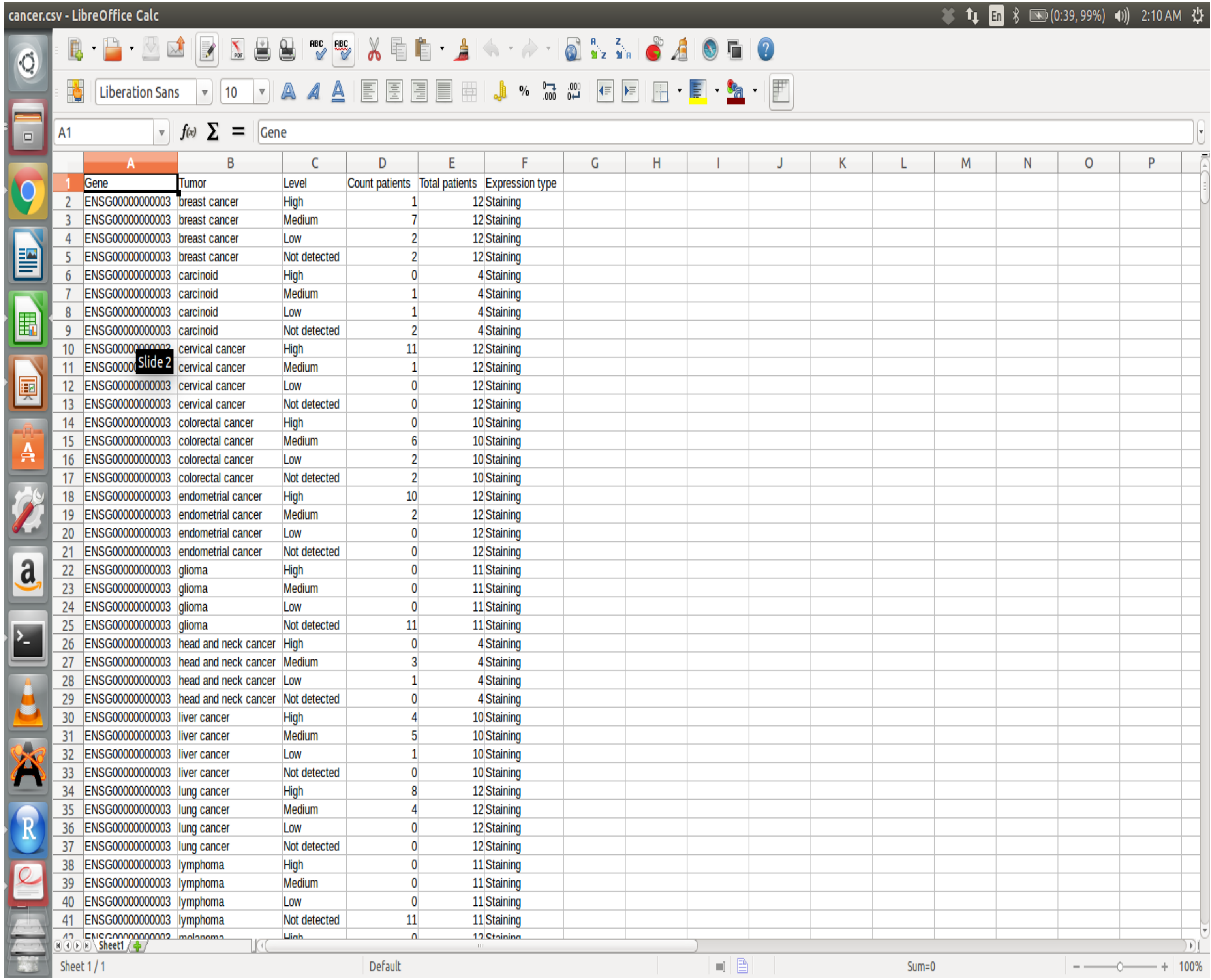
COMPARATIVE ANALYSIS ACROSS CANCERS USING PROTEIN DATA

Noor Pratap Singh

DATA

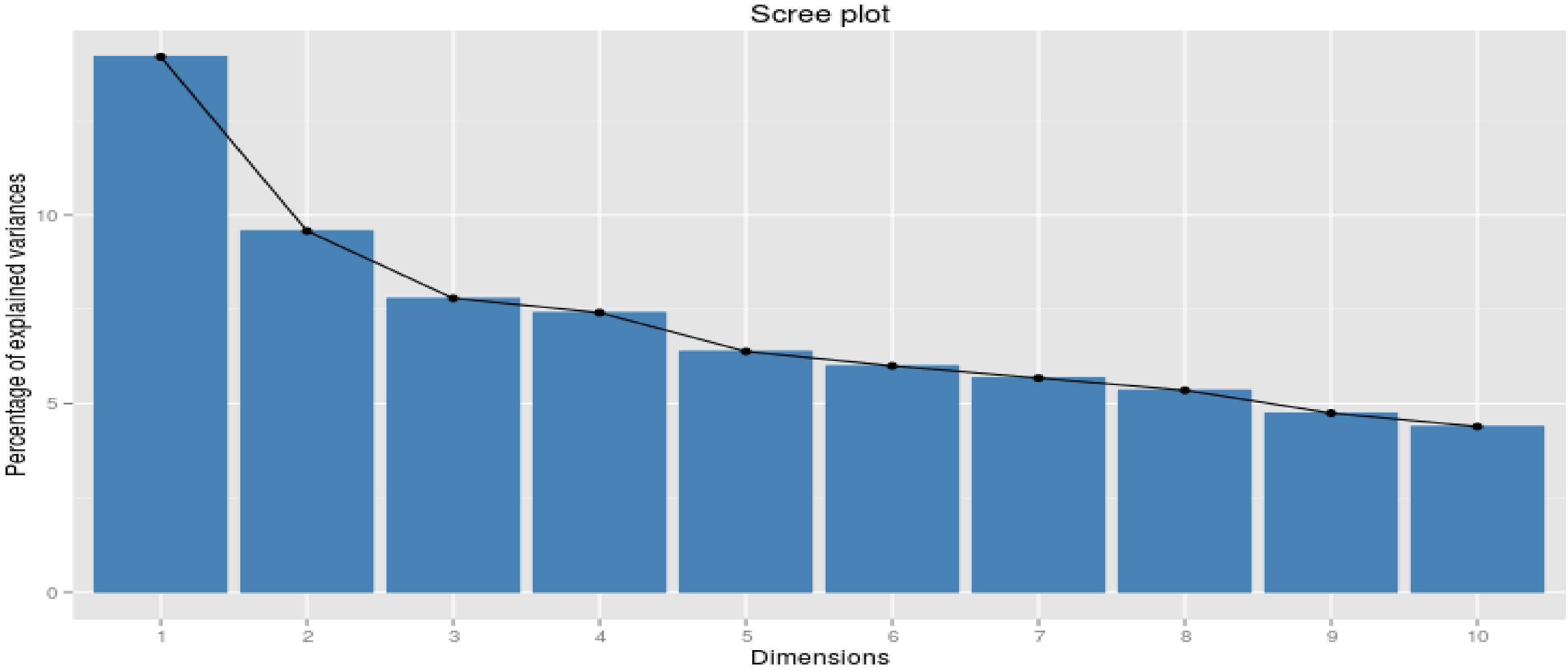
- We worked with cancer data and tissue data.
- Taken from Human Protein Atlas(HPA).
- Cancer data contains a total of 16613 unique Ensembl Ids.
- We also took specific patients' data from http://msb.embopress.org/content/10/3/721



	A	B	C	D	E	F
1	Gene	Tumor	Level	Count patients	Total patients	Expression type
2	ENSG00000000003	breast cancer	High	1	12	Staining
3	ENSG00000000003	breast cancer	Medium	7	12	Staining
4	ENSG00000000003	breast cancer	Low	2	12	Staining
5	ENSG00000000003	breast cancer	Not detected	2	12	Staining
6	ENSG00000000003	carcinoid	High	0	4	Staining
7	ENSG00000000003	carcinoid	Medium	1	4	Staining
8	ENSG00000000003	carcinoid	Low	1	4	Staining
9	ENSG00000000003	carcinoid	Not detected	2	4	Staining
10	ENSG00000000003	cervical cancer	High	11	12	Staining
11	ENSG0000 Slide 2	cervical cancer	Medium	1	12	Staining
12	ENSG00000000003	cervical cancer	Low	0	12	Staining
13	ENSG00000000003	cervical cancer	Not detected	0	12	Staining
14	ENSG00000000003	colorectal cancer	High	0	10	Staining
15	ENSG00000000003	colorectal cancer	Medium	6	10	Staining
16	ENSG00000000003	colorectal cancer	Low	2	10	Staining
17	ENSG00000000003	colorectal cancer	Not detected	2	10	Staining
18	ENSG00000000003	endometrial cancer	High	10	12	Staining
19	ENSG00000000003	endometrial cancer	Medium	2	12	Staining
20	ENSG00000000003	endometrial cancer	Low	0	12	Staining
21	ENSG00000000003	endometrial cancer	Not detected	0	12	Staining
22	ENSG00000000003	glioma	High	0	11	Staining
23	ENSG00000000003	glioma	Medium	0	11	Staining
24	ENSG00000000003	glioma	Low	0	11	Staining
25	ENSG00000000003	glioma	Not detected	11	11	Staining
26	ENSG00000000003	head and neck cancer	High	0	4	Staining
27	ENSG00000000003	head and neck cancer	Medium	3	4	Staining
28	ENSG00000000003	head and neck cancer	Low	1	4	Staining
29	ENSG00000000003	head and neck cancer	Not detected	0	4	Staining
30	ENSG00000000003	liver cancer	High	4	10	Staining
31	ENSG00000000003	liver cancer	Medium	5	10	Staining
32	ENSG00000000003	liver cancer	Low	1	10	Staining
33	ENSG00000000003	liver cancer	Not detected	0	10	Staining
34	ENSG00000000003	lung cancer	High	8	12	Staining
35	ENSG00000000003	lung cancer	Medium	4	12	Staining
36	ENSG00000000003	lung cancer	Low	0	12	Staining
37	ENSG00000000003	lung cancer	Not detected	0	12	Staining
38	ENSG00000000003	lymphoma	High	0	11	Staining
39	ENSG00000000003	lymphoma	Medium	0	11	Staining
40	ENSG00000000003	lymphoma	Low	0	11	Staining
41	ENSG00000000003	lymphoma	Not detected	11	11	Staining
42	ENSG00000000003	melanoma	High	0	12	Staining

MCA Analysis

- We did MCA Analysis on the entire set of genes.

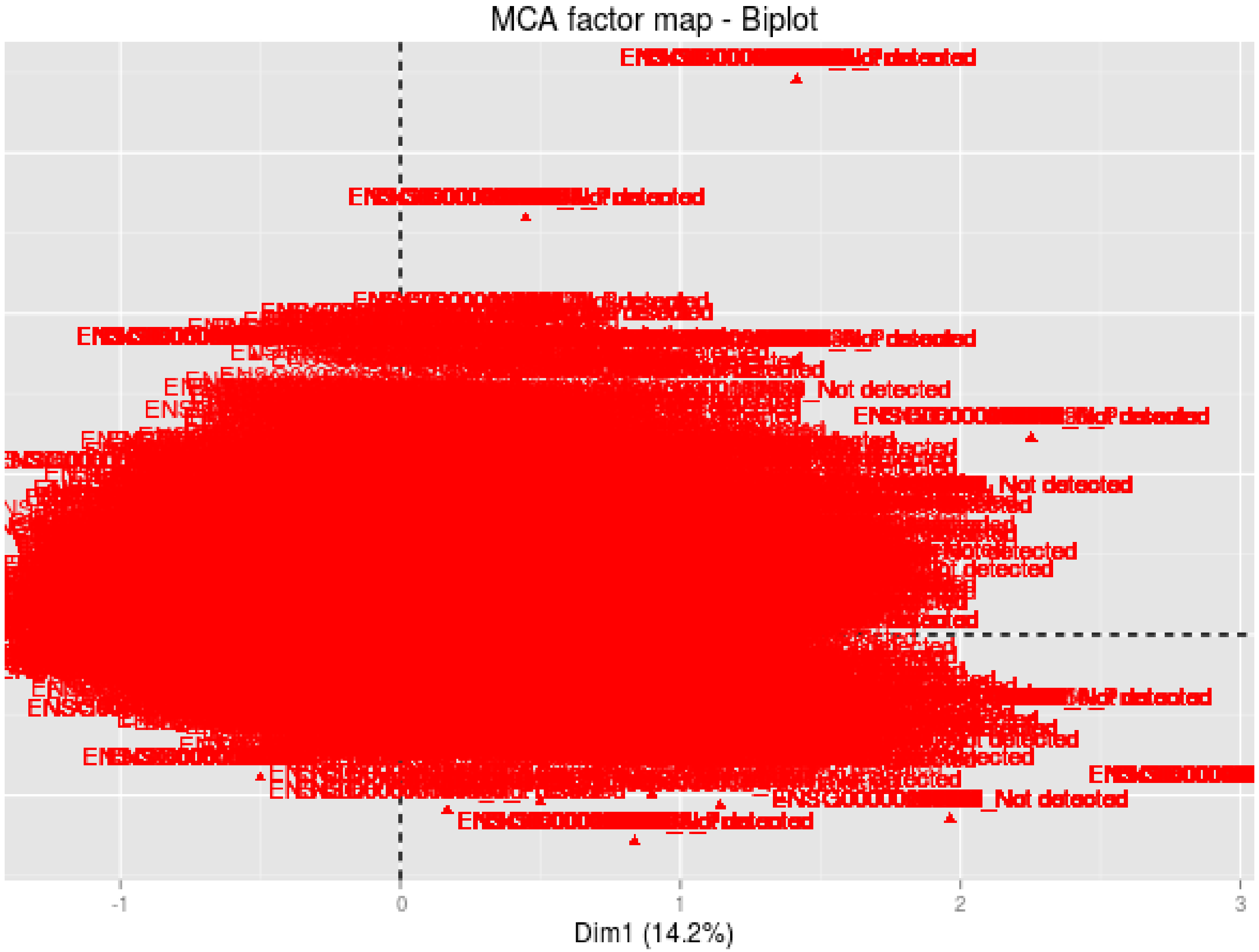


We get the above contribution along the respective dimensions.

Overall MCA Plot

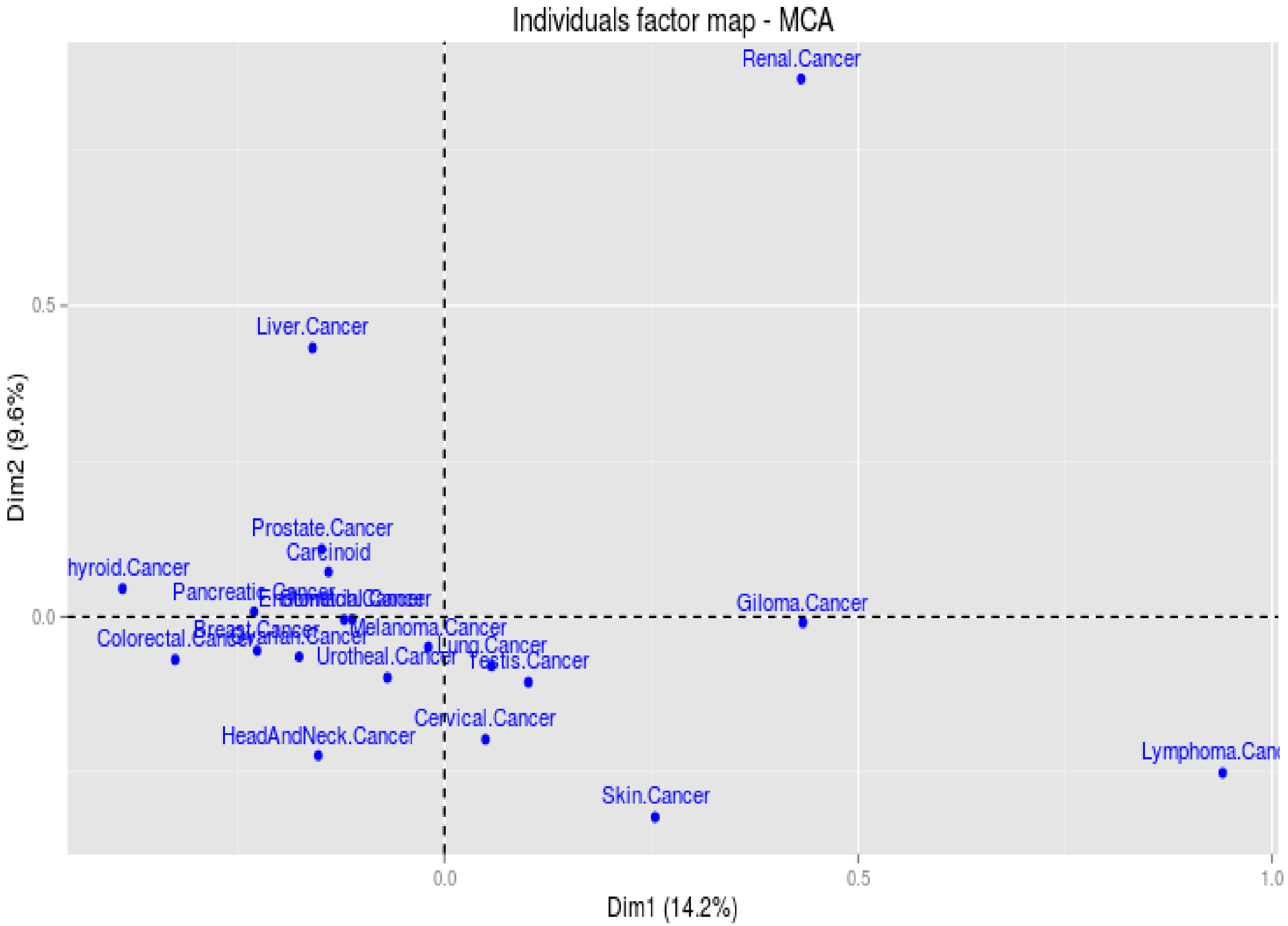
The picture is clumsy and not much can be interpreted from it.

But it plots the entire variables as well as the individuals and the variables being huge messes it up.

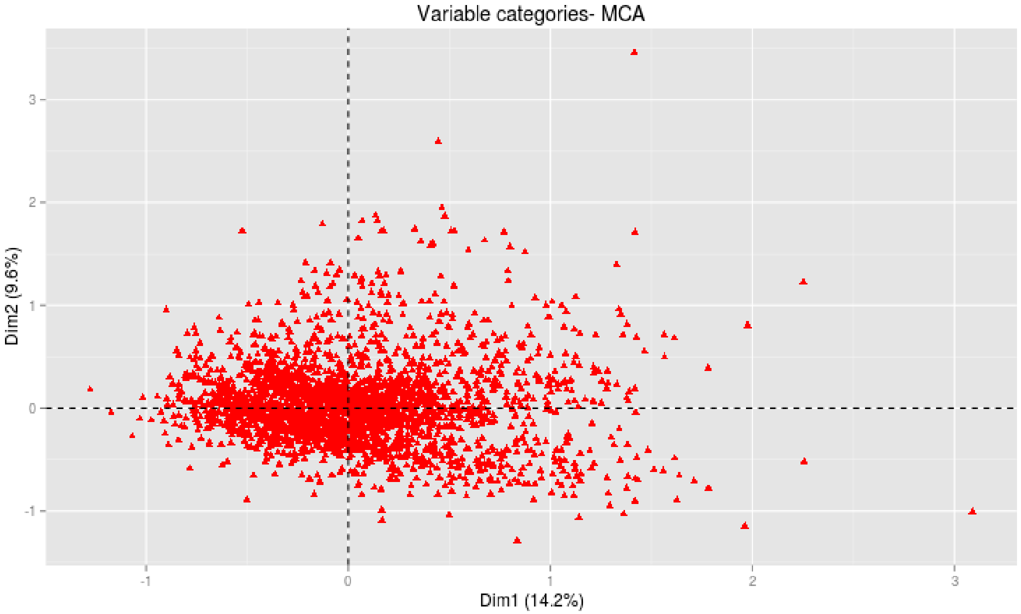


Some Cancers are clearly separated !

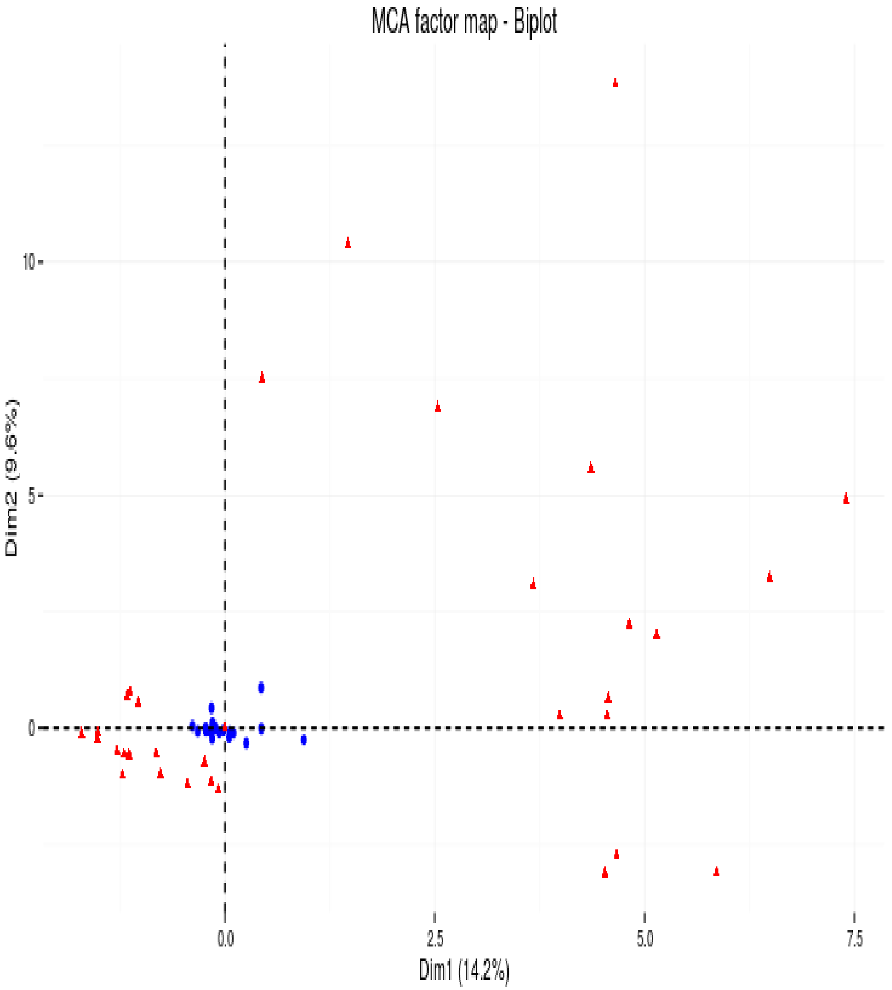
- We plot an individual wise plot.
- We can clearly see not all cancers have uniform distribution with respect to gene levels.



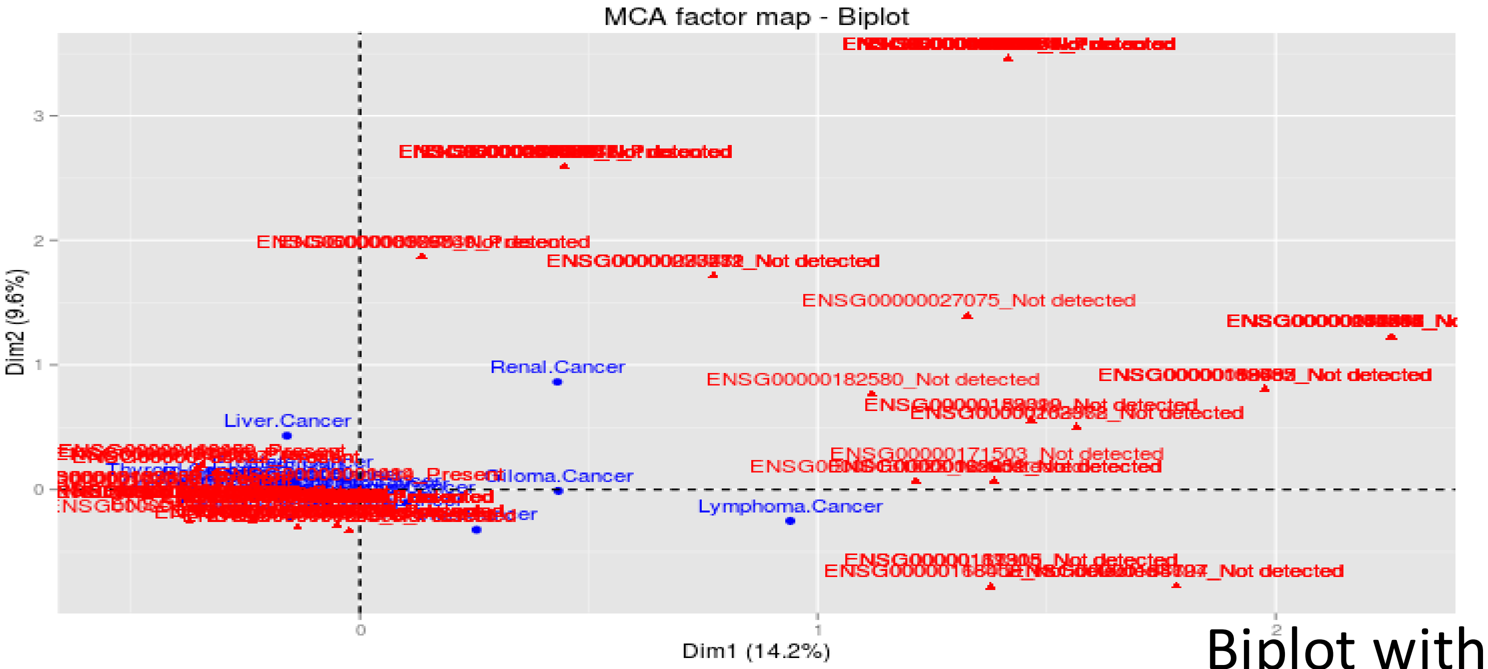
Other Plots



All variables



Asymmetric Plot

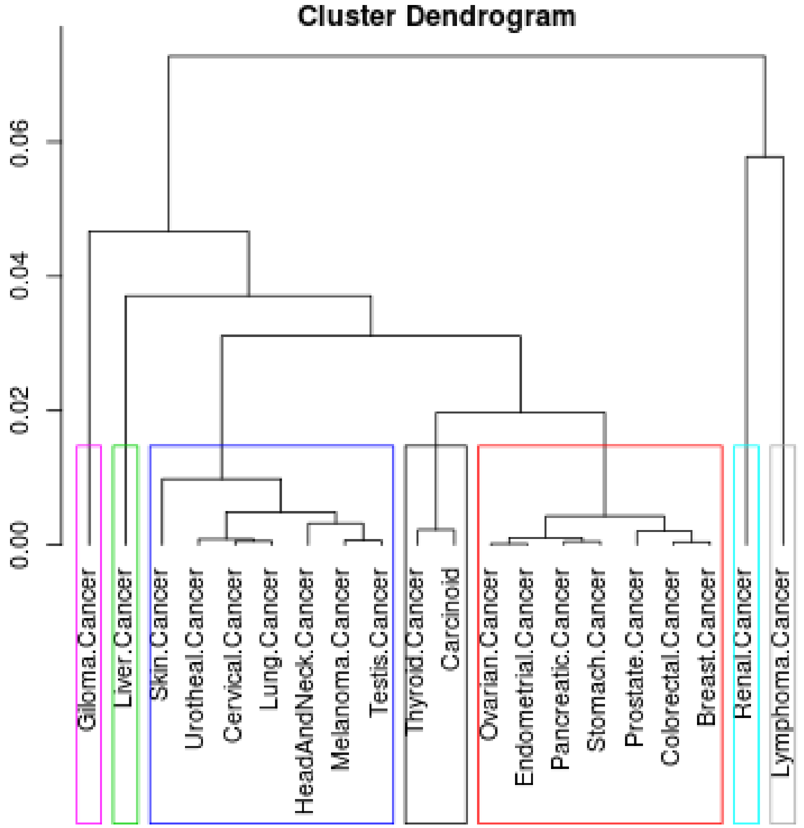


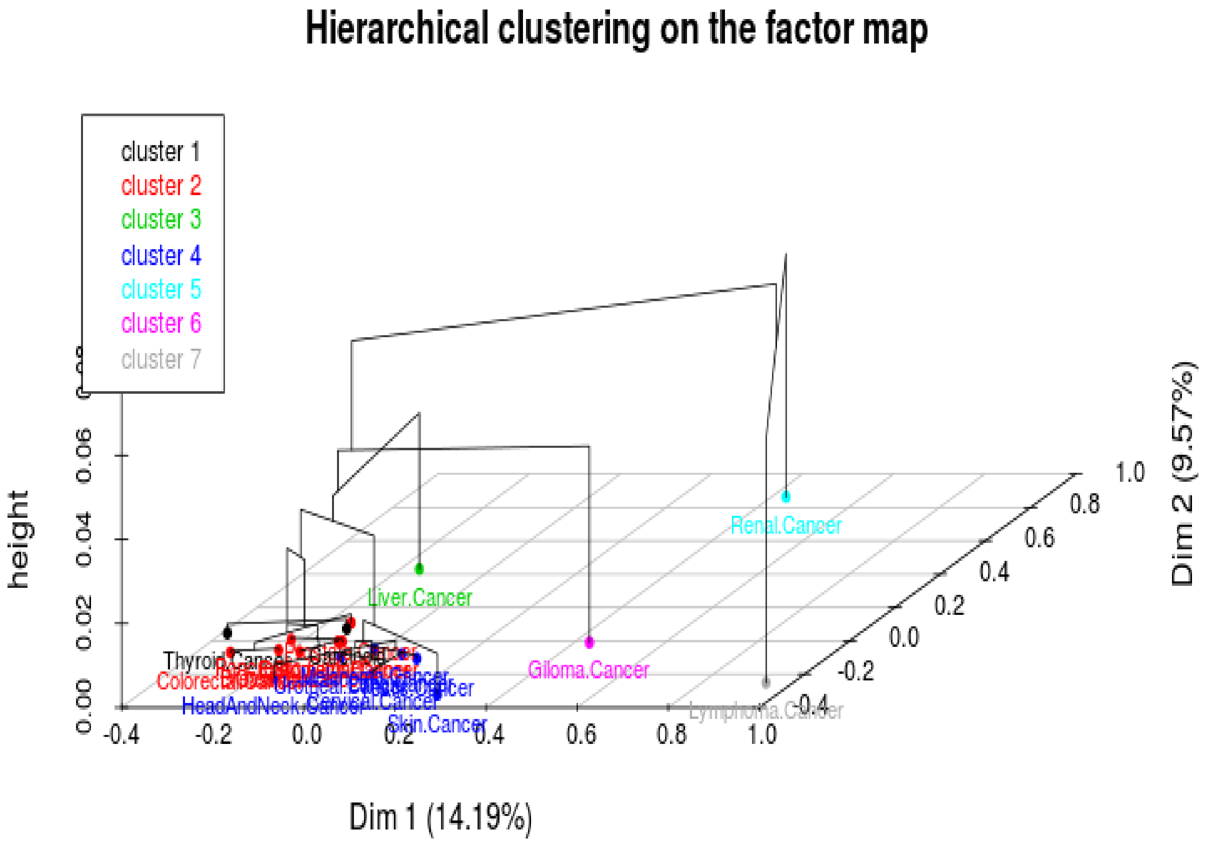
Biplot with cos2 > 0.6

Cluster Analyis

- We still don't know the exact dimensions to chose.
- So we do cluster analysis to actually find out which cancers can be actually clubed together.
- We get a total of 7 clusters.

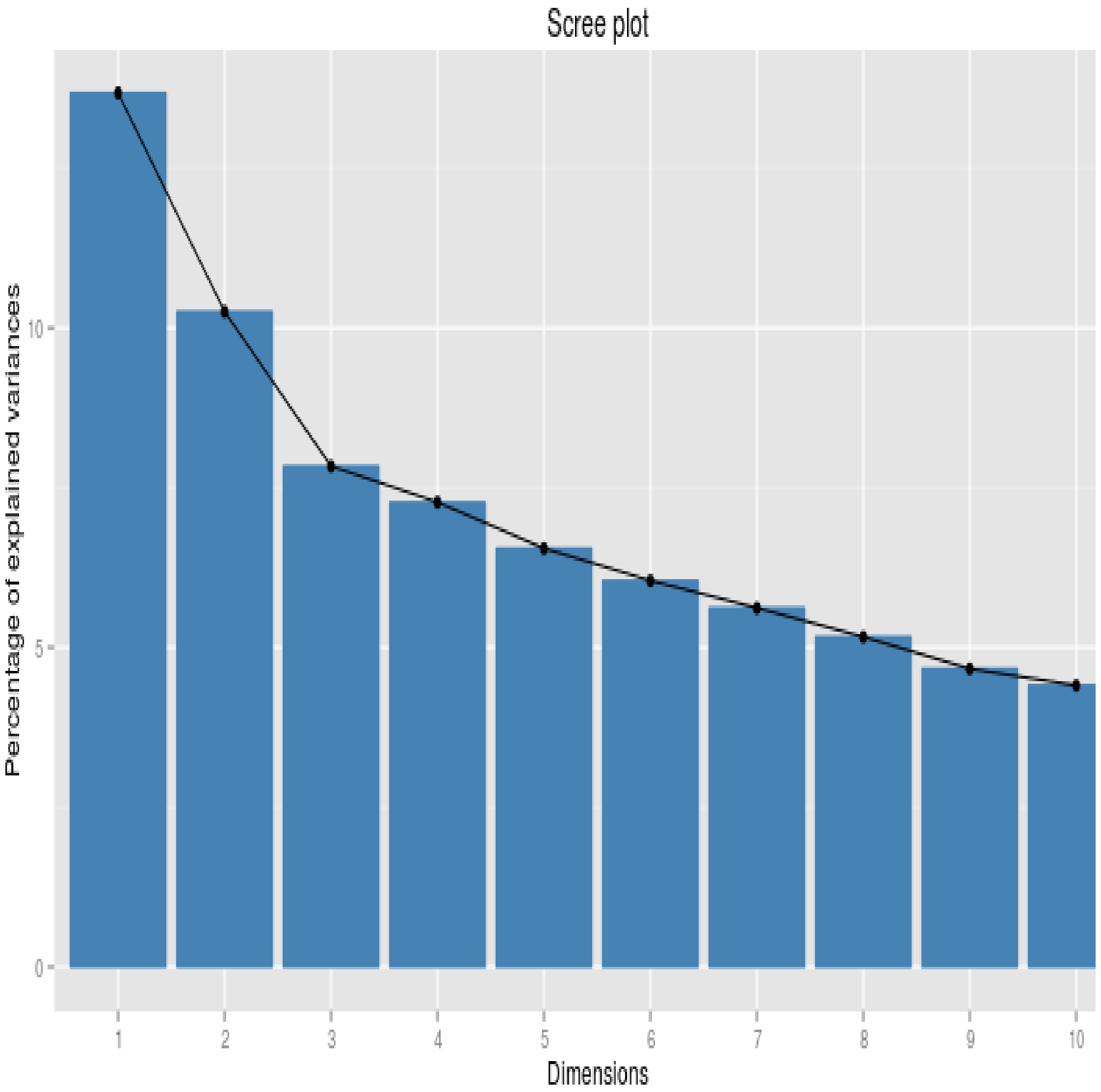
Clustering

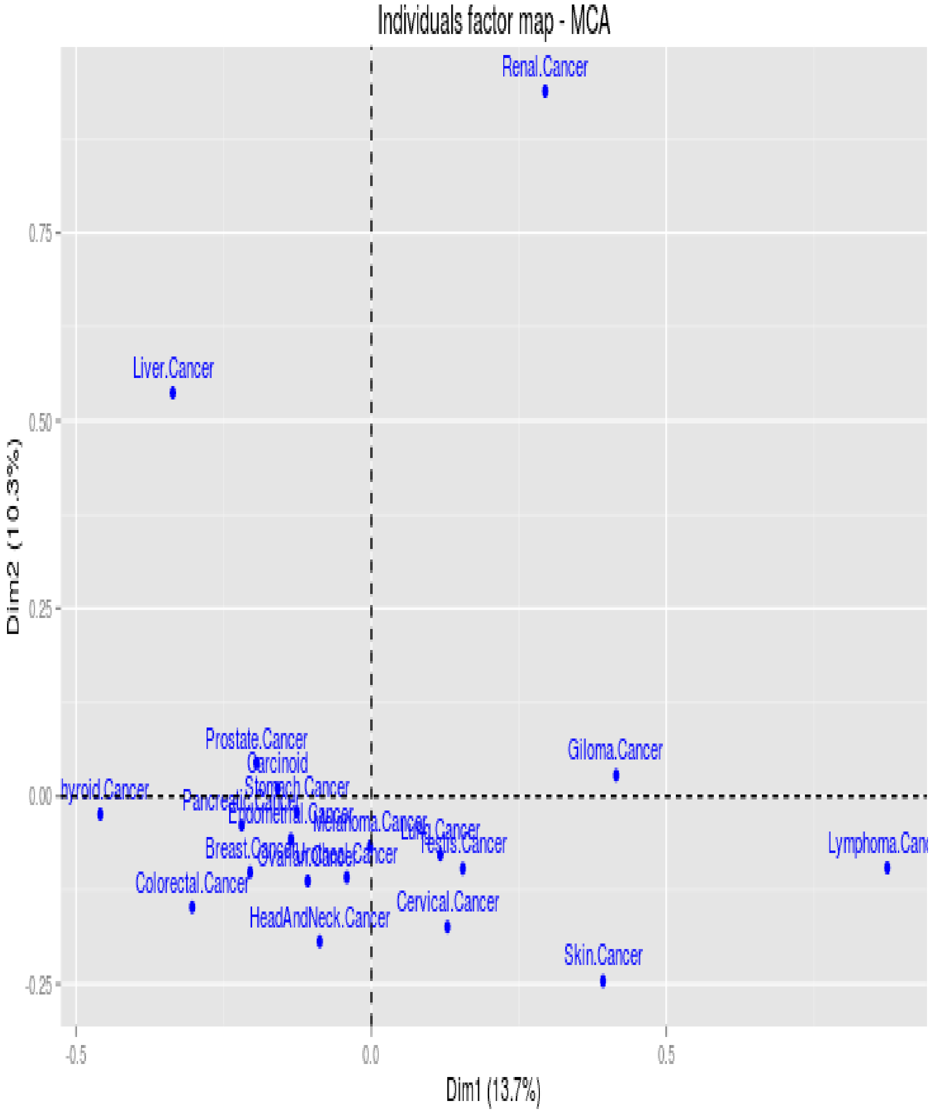




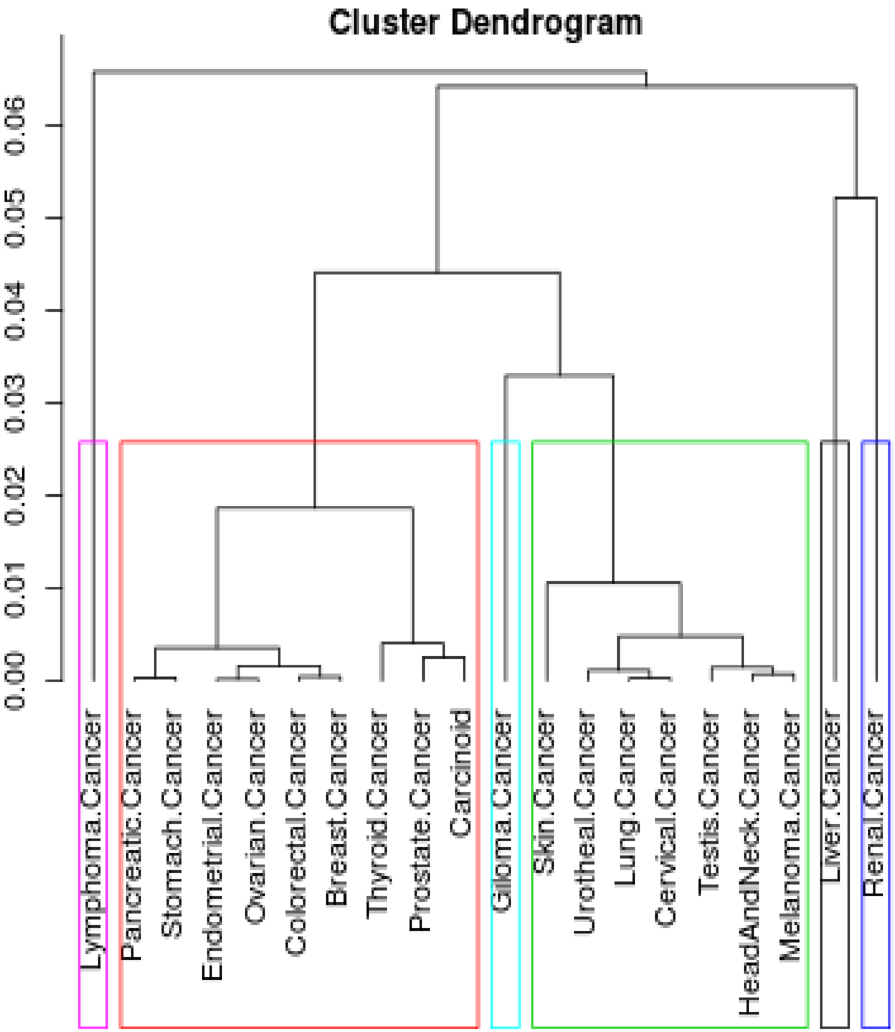
MCA for metabolic genes

- We do the same analysis on the same data but only for metabolic genes.

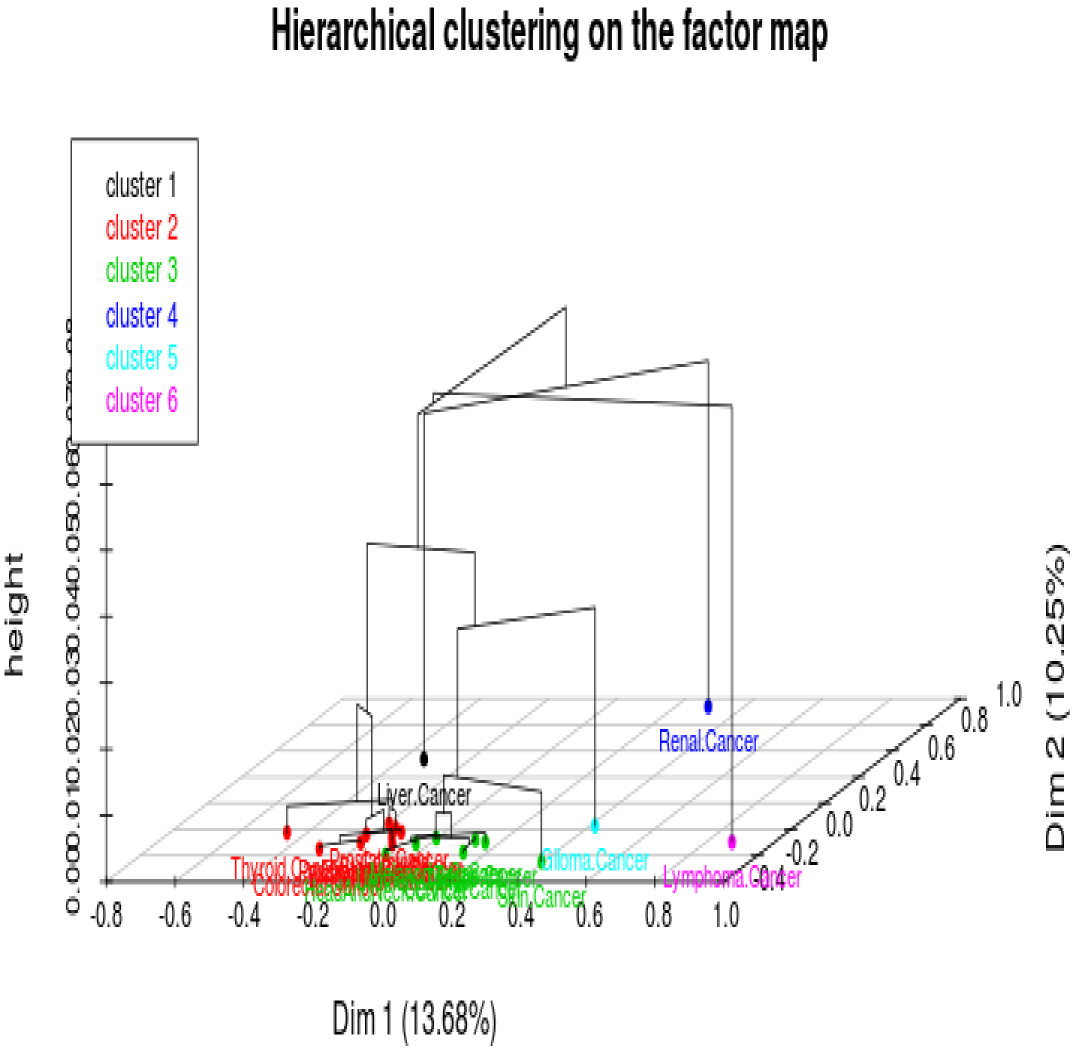




Cluster Analysis



Cluster Dendrogram



Hierarchical clustering on the factor map

- Overall individual wise both look very similar.
- The cluster analysis brings a different story.
- There are 7 clusters in former while 6 clusters in latter.
- Also position of cancers in cluster change.
- So far we have considered only metabolic genes.
- We plan to perform similar analysis based on other proccesess such as cell cycle etc.

Hepatocellular Carcinoma

Comparing Normal and Cancer Tissue data

Problems

- Cancer data associated with each gene for each cancer 4 levels and their counts.
- Normal tissue data on the other hand had associated with it each a cell type with only *one* level and no count.
- So it made comparing the two datasets difficult.

Solution

- So in order to fix a level with respect to gene for a cancer we make 4 basis of comparison(50,75,90, 100) based on count clubbing the 4 levels into 2.

Identification of Differentially Expressed Genes(DEG)

- Once the level of liver cancer w.r.t to each gene is calculated based on type of mode(50,75,90,100) we compare it with normal tissue data of hepatocytes.
- The genes which do not match contribute to the differentially expressed genes(DEG).
- As per our calculation we get the following number of DEG:
  1. **50% -** 3547
  2. **75% -** 1532
  3. **90% -** 697
  4. **100%-** 310

Functional Annotation

- After identification of DEGs for functional annotation of DEGS we tried different tools like David, Genecodis and GOSim package in R.
- For DAVID analysis we used BP 5 and we get following results:
  - **50%** 84 GoTerms with 34 having p value < 0.05
  - **75%** 53 GoTerms with 25 having p value < 0.05
  - **90%** 18 GoTerms with 9 having p value < 0.05
  - **100%** 4 GoTerms with 1 having p value < 0.05
- The results were not very satisfactory.

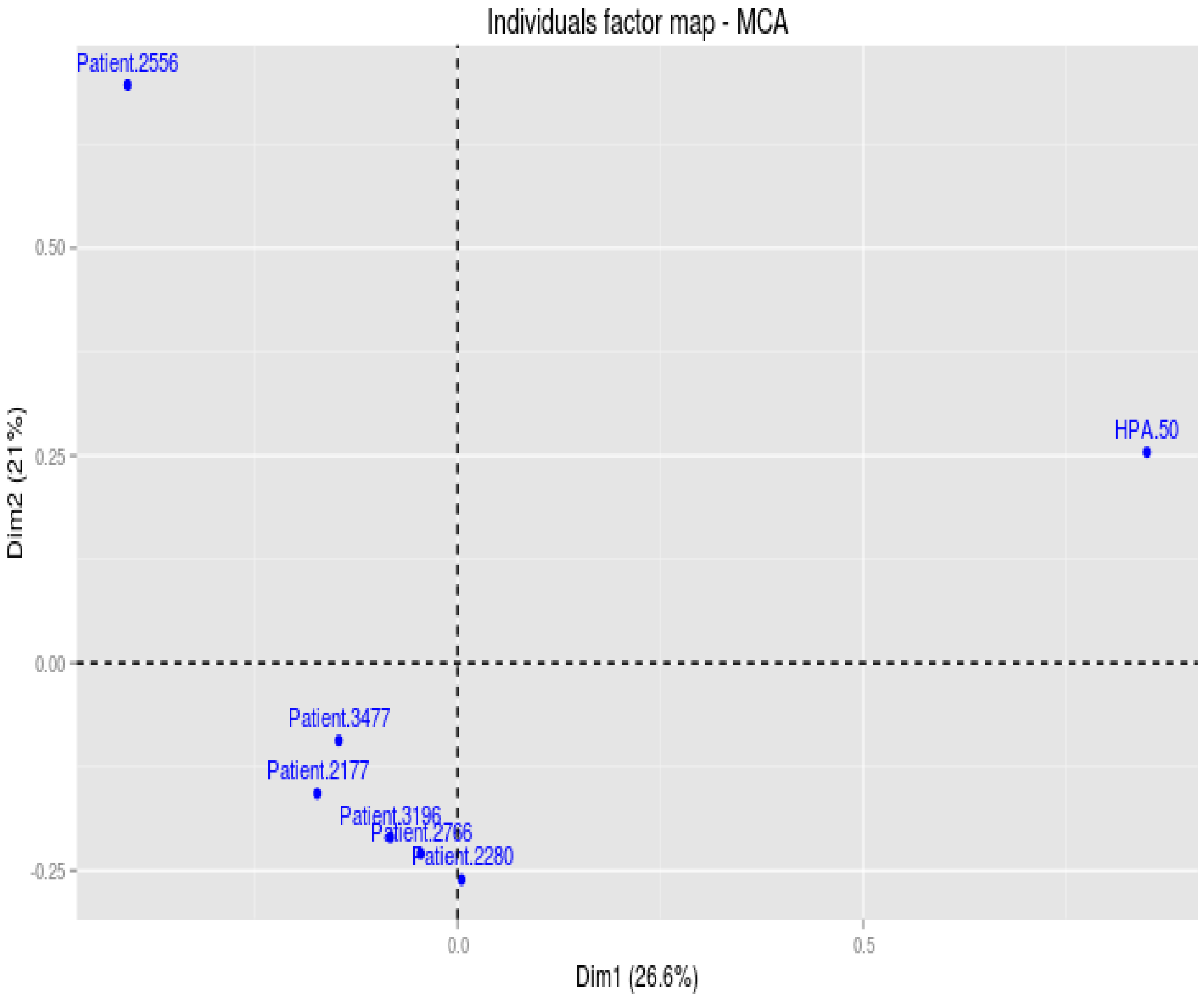
GOSim

- We used 75 % differentially expressed genes.
- We did GOenrichment analysis on the above set.
- We get a total of 32 different GOTerms which do not match DAVID.
- We also found the similarity matrix using Jiang and Conrath method.
- Based on above similarity matrix we use k mediod clustering with k = 8 (maximum silhouette width tried for different values of k)
- We again did GOenrichment analysis on individual clusters.

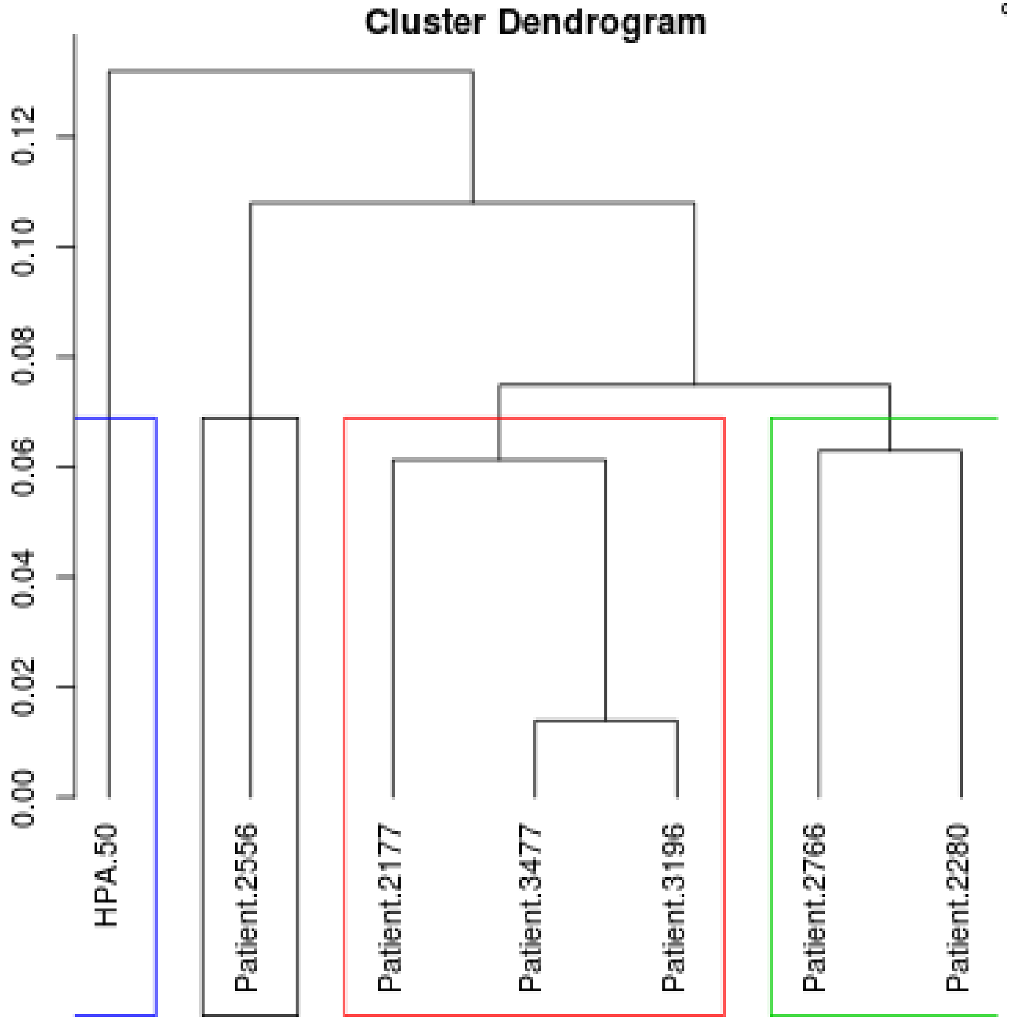
Patient Wise Comparison

- In order to verify that whether we have correctly picked DEGs we also take patient data.(The number of genes are less than HPA and 6 patients)
- We also do a patient wise protein expression comparison with HPA at 50% since it is the most generic.
- We get the following number of proteins with same expression level across patients: 11946, 12414, 11442, 12206, 12258, 12151.
- A question might arise that IDs that do not match may be the DEGS but if you match DEGS at 50% with above genes you get the following:
  - 813, 1198, 666, 1130, 979, 856.
- A point worth noting that 50 % is the most generic case and the patients might be in different stages of cancer.

MCA Analysis



Clustering


Cluster Dendrogram
HPA.50
Patient.2556
Patient.2177
Patient.3477
Patient.3196
Patient.2766
Patient.2280