Diagnosing Cancer using Gene Signatures

Jared Gaertner*
UVic CSc Undergraduate

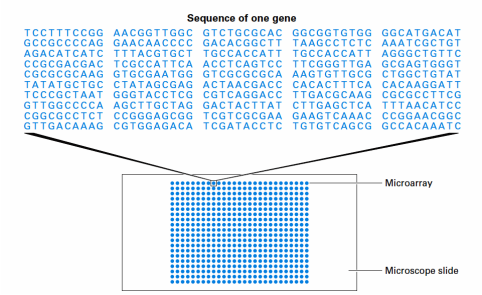

Figure 1: *Microarray [Campbell and Heyer 2006]*

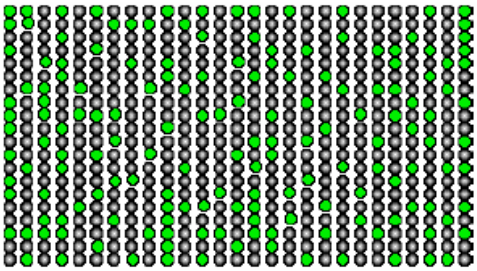

Figure 2: *microarray green [Campbell and Heyer 2006]*

1 Introduction

The increasing knowledge and technical advances in Bioinformatics has allowed a new area to open up, which is diagnosing cancer using gene signatures. It is a relatively new field, which has been brought on by the advances in DNA expression analysis, and in particular, DNA microarrays.

2 DNA Microarrays

DNA microarrays allow for all genes of a genome to be measured for environmental response on one slide. An example of a DNA microarray can be seen in Fig. 1. All the genes are amplified by polymerase chain reaction (PCR). Each spot indicates an amplified copy of a single gene and has an actual diameter of about 100 μm.

The main advantage of the microarray is the ability to have a visual interpretation of an entire genome all on one slide, which can then be analyzed to take in data. The principle behind microarrays is that the cells of the organism, which is to be observed, will react differently to different environments. The different environments are generally the presence or the lack of oxygen. These cells are then grown in each of these environments, and the mRNA from the cell is obtained and converted into complementary DNA (cDNA). These cDNA contain different dyes in order to determine which environment the cDNA were grown in. The typical dyes used are green (Cy3) and red (Cy5). These cDNA are then mixed and incubated with the microarrays and observed over that time, each of the different dyes on different slides. Due to the pigments, the various dots (genes) on the microarray will begin to become more coloured if that gene is being expressed more. After these slides have been allowed to incubate (and at time intervals during the incubation), the fluorescence is measured for both the red and green slides. A simplified example of the red and green slides can be seen in Fig. 3 and Fig. 2 respectively. The combined fluorescence of both slides is obtained, giving a yellow output, as seen in Fig. 4.

Those figures are merely demonstrative in order to give a more clear cut example of the colouring scheme. In reality, the colours are not binary in fashion, instead they range from black to either green or red depending on the slide, and can be an infinite number of possibilities. An example of a real-life reading of the fluorescence of a microarray can be seen in Fig. 5.

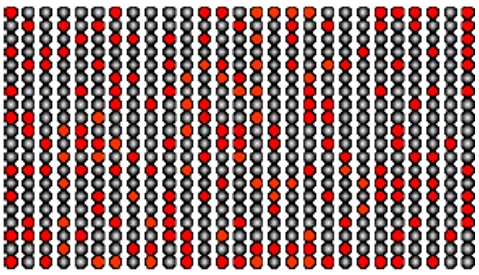

Figure 3: *Microarray red [Campbell and Heyer 2006]*

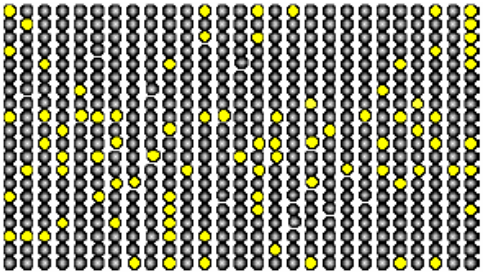

Figure 4: *Microarray yellow [Campbell and Heyer 2006]*

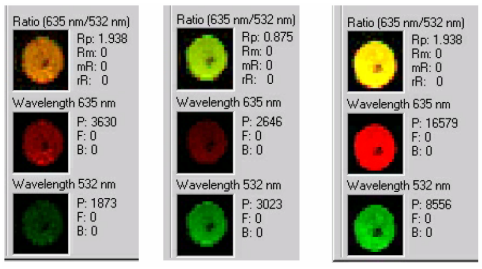

Figure 5: *microarray actual [Campbell and Heyer 2006]*

*e-mail: jaredg@uvic.ca

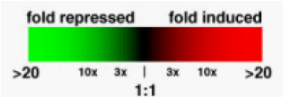

Figure 6: *Microarray colour scale [Campbell and Heyer 2006]*

The data is then obtained, however, it is just a large array of fluorescences, which is not useful to the person who will be interpreting the data. In order to have the data in a useful manner, it must be visualized in a way that the patterns of the gene expression will become evident when they are sorted or compared to similar genes. The presence of oxygen is the control condition and the absence is the experimental condition. If the gene is induced in the experimental condition, it is represented by a proportional red colour. If the gene is repressed in the experimental condition, it is represented by a proportional green colour. This ratio of red to green (also known as the fold change), is in fact just the ratio of the experimental to the control condition, which is a good indicator of the gene expression. This ratio is then applied to a scale, as seen in Fig. 6.

An example of the output of the microarray process can be seen in Fig. 7. The block, row, and column data are related to the location of the gene on the microarray slide, and as yeast has a smaller genome, it is a good example to examine, as all of it genes have been identified. However, even though the genes of yeast have been identified, it is still too large to show an example of have the microarray data is analyzed. For this, a set of hypothetical data can be seen in Fig. 8. The first thing that can be seen, is that there is data at various times during the incubation. This would be a problem for the visualization, as it would become to convoluted to display different visualizations at different time steps. Instead, the data is used in order to find one final value which can be used for a particular gene. The way this is done is through the Pearson correlation coefficient(r). This allows each gene to be compared to the other genes in order to find how similar they are. The Pearson correlation coefficient, for a gene A to gene B, is found as follows

$$\overline{x}_A = \frac{1}{n}\sum_{i=1}^{n} f_i \tag{1}$$

$$s_A = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_A)^2} \tag{2}$$

$$\text{Norm}_A = \{\frac{f_1 - \overline{x}_A}{s_A}, \frac{f_2 - \overline{x}_A}{s_A}, \cdots, \frac{f_n - \overline{x}_A}{s_A}\} \tag{3}$$

$$r = \sum_{i=1}^{n} \text{Norm}_{A,i} \cdot \text{Norm}_{B,i} \tag{4}$$

where f is the fold change, and n is the total number of data points taken for a gene. This must also be completed on every other gene pairing. With the given data in 8, a problem with the correlation values can appear. However, this method can cause problems due to the fact that correlation is sensitive to the magnitude of the patterns, so generally the values are first log-transformed. The output of the correlation values for a few of the gene pairs, which were first log-transformed before finding the correlation values, are shown in Fig. 9. Now this data is in a form which shows the pairwise distance between the various genes, which allows for clustering to be done on the data.

Block	Column	Row	Gene Name	Red	Green	Red:Green Ratio
1	1	1	*tub1*	2345	2467	0.95
1	1	2	*tub2*	3589	2158	1.66
1	1	3	*sec1*	4109	1469	2.80
1	1	4	*sec2*	1500	3589	0.42
1	1	5	*sec3*	1246	1258	0.99
1	1	6	*act1*	1937	2104	0.92
1	1	7	*act2*	2561	1562	1.64
1	1	8	*fus1*	2962	3012	0.98
1	1	9	*idp2*	3585	1209	2.97
1	1	10	*idp1*	2796	1005	2.78
1	1	11	*idh1*	2170	4245	0.51
1	1	12	*idh2*	1896	2996	0.63
1	1	13	*erd1*	1023	3354	0.31
1	1	14	*erd2*	1698	2896	0.59

Figure 7: *Example data for one yeast DNA microarray*

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene C	1	8	12	16	12	8
gene D	1	3	4	4	3	2
gene E	1	4	8	8	8	8
gene F	1	1	1	0.25	0.25	0.1
gene G	1	2	3	4	3	2
gene H	1	0.5	0.33	0.25	0.33	0.5
gene I	1	4	8	4	1	0.5
gene J	1	2	1	2	1	2
gene K	1	1	1	1	3	3
gene L	1	2	3	4	3	2
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125

Figure 8: *Hypothetical data for 12 genes fold change*

	gene C	gene D	gene E	gene F	gene G	gene H	...
gene C	1						
gene D	0.94	1					
gene E	0.96	0.84	1				
gene F	-0.40	-0.10	-0.57	1			
gene G	0.95	0.94	0.89	-0.35	1		
gene H	-0.95	-0.94	-0.89	0.35	-1	1	
...							...

Figure 9: *Gene pair correlation*

3 Clustering

Given data similar to Fig. 9, clustering can be completed. Various types of clustering can be performed, such a k-means clustering, however, hierarchical clustering is the most popular option. The algorithm for hierarchical clustering can be seen in *Hierarchical-Clustering*.

Algorithm *HierarchicalClustering*
Input: Pairwise correlation matrix
Output: Hierarchical tree
1. Let each data point in M be a cluster
2. Compute the distance matrix
3. **repeat**
4. Merge the two closest clusters
5. Update the distance matrix
6. **until** Only one cluster remains
7. **return** Hierarchical tree

The main point of the algorithm, is how the distances are computed. A common way this is done is called Unweighted Pair-Group Method using an arithmetic Average (UPGMA), and is defined as follows:

$$\text{distance}(C_i, C_j) = \frac{\sum_{\forall p_x \in C_i} \sum_{\forall p_y \in C_j} \text{distance}(p_x, p_y)}{|C_i||C_j|} \quad (5)$$

where C denotes a cluster, p denotes a gene, distance(p_x, p_y) finds the pairwise distance between gene x and y, and distance(C_i, C_j) finds the distance between clusters i and j.

4 Gene Signatures

The results of the microarrays, colouring scheme, and clustered data can find signature genes. Multiple cells can be individually clustered, then those cells can also be clustered to one another based on those found gene expression profiles. The result of this is signature genes. Signature genes are genes which can be used to describe a particular cell type. Diffuse Large B-cell Lymphoma (DLBCL), an aggressive form of cancer, was taken through the process of creating its microarrays from various cancerous cells and non-cancerous cells, 96 cells in total. The pairwise matrices were determined from the microarray data, and the results were clustered, as shown in Fig. 10. The gene signatures can be seen on the right of Fig. 10, which are matched from the hierarchical clustering completed and known gene expression data.

The DLBCL samples and two germinal center (GC) B-cell lines were reclustered using the GC signature genes found in Fig. 10, with about half the cells having activated B-like qualities in blue, with the other half having GC B-like qualities in orange. The results are shown in Fig. 11*a*. The original GC B-cells are shown in black. Reclustering was then completed on all of the genes based on the GC/Activated B-like DLBCL (orange/blue) classification, shown in Fig. 11*b*. Fig. 11*c* shows the zoomed in clustering on just the GC and activated B-cell portion. Fig. 11*d* shows the names of the individual cells on the hierarchical tree found by the clustering.

This classification of DLBCL cells into germinal center B-like DLBCL and activated B-like DLBCL allowed for a clear-cut difference between the cancer cells of patients. This data was then observed with regards to patients and whether their cancer cells were GC or activated B-like DLBCL. The survival rates for DLBCL versus the type of classification of their DLBCL cells is shown in Fig. 12*a*. This shows a very drastic difference in survival rates of GC B-like versus activated B-like DLBCL cells, with the patients of GC B-like

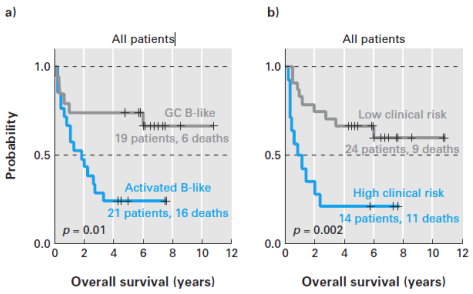

Figure 12: *DLBCL survival rates [Campbell and Heyer 2006]*

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene D	1	3	4	4	3	2
gene L	1	2	3	4	3	2

Figure 13: *Correlation example*

DLBCL cells having a much higher survival rate. This was compared to patients separated using IPI indices (a traditional clinical criteria), shown in Fig. 12*b*, which did not give quite as good of results as the GC B-like versus activated B-like DLBCL cells classification. This shows that the gene signatures give a more accurate test for survival rates, which is very important factor into whether or not chemotherapy should be considered, as it is a very difficult process which some patients may not want to go through if they have a low survival rate.

5 Example

An example of the process of using gene signatures to diagnosing cancer will be performed on the data from 8. This is due to the fact that it is an easier example to do to completion, as well as the fact that pairwise data is not readily available for clustering. The first factor to determine with this hypothetical data, is the Pearson correlation factor. For the data from Fig. 13, the correlation factor is found as follows (for data which is not first log-transformed)

$$\overline{x}_D \approx 2.83, \overline{x}_L = 2.5$$

$$s_D \approx 1.067, s_L \approx 0.957$$

$$\text{Norm}_D = \{-1.715, 0.1593, 1.097, 1.097, 0.1593, -0.7779\}$$

$$\text{Norm}_L = \{-1.567, -0.5225, -.5225, 1.567, 0.5225, -0.5225\}$$

$$r_{D,L} = 0.897$$

The data from the correlation values for all the pairs can then be seen in Fig. 9, which has used data that was first log-transformed unlike the previous equations completed. The clustering steps can be seen can be seen in Fig. 14, and the visual output is in Fig. 15.

If this was real data, that visualization and data could then be compared to known gene expressions in order to determine the location of those such genes, as well as to determine the differences between the cells which have just been clustered in order to determine different classifications of those cells.

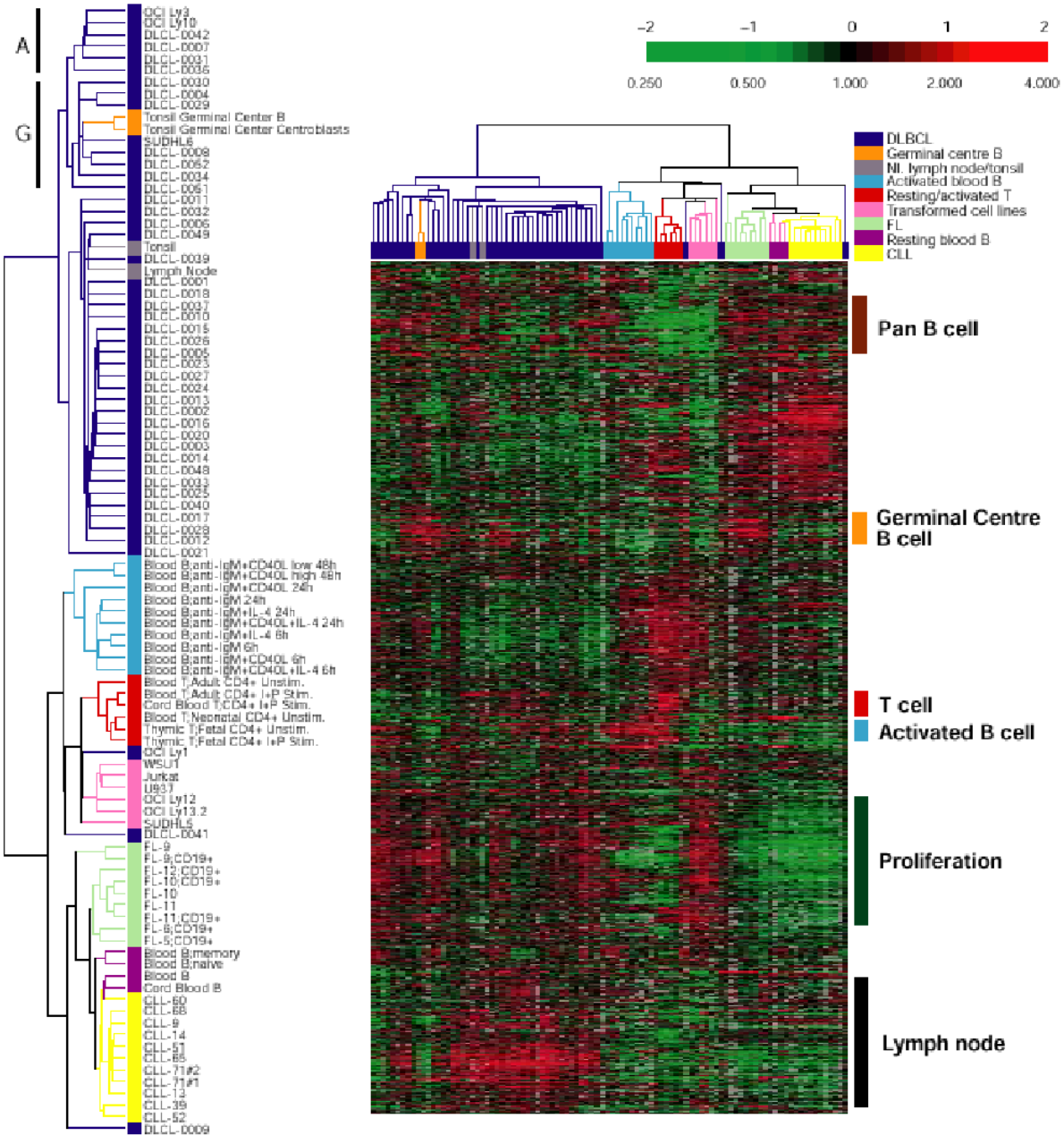

Figure 10: *DLBCL gene expression [Campbell and Heyer 2006]*

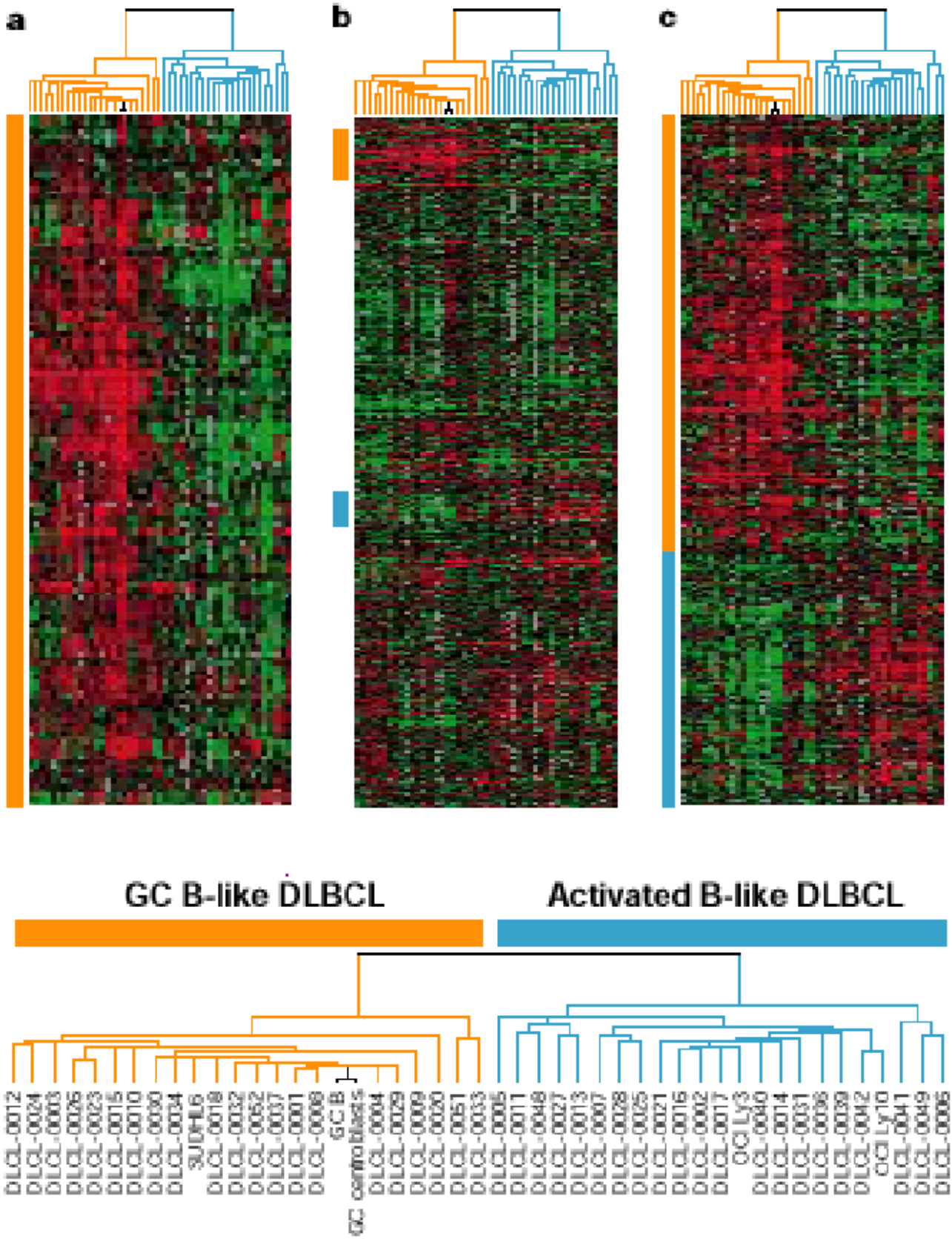

Figure 11: *DLBCL gene expression signatures [Campbell and Heyer 2006]*

#	Objects		r	New Object
1	L	G	1.00	[LG]
2	E	C	0.96	[EC]
3	N	H	0.95	[NH]
4	M	[NG]	0.95	[MNH]
5	[LG]	D	0.94	[LGD]
6	[EC]	[LGD]	0.94	[ECLGD]
7	I	F	0.60	[IF]
8	J	[ECLGD]	0.29	[JECLGD]
9	K	[JECLGD]	0.19	[KJECLGD]
10	[KJECLGD]	[IF]	-0.12	[KJECLGDIF]
11	[MNH]	[KJECLGDIF]	-0.96	[MNHKJECLGDIF]

Figure 14: *Hierarchical clustering example*

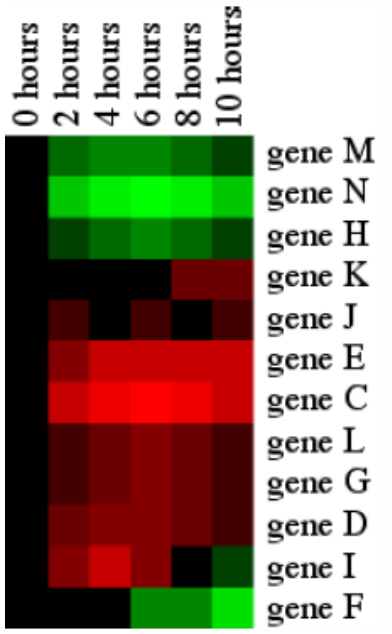

Figure 15: *Cluster results [Campbell and Heyer 2006]*

6 Conclusion

The improvement in technology in the areas of Bioinformatics can greatly affect the lives of many, especially with such a difficult disease such as cancer. Providing various ways to classify cancer patients in order to give them more information and better insight into the future challenges is a important task. Microarrays provide a much quicker and more effective way to identify gene signatures for an entire genome than previous applications. This allows for quicker evaluation of the genes and more accurate information.

References

CAMPBELL, A. M., AND HEYER, L. J. 2006. *Discovering Genomics, Proteomics and Bioinformatics*, 2 ed. Cold Spring Harbor Laboratory Press and Benjamin Cummings, February. ISBN 0-8053-4722-4.