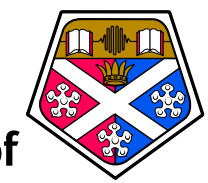
NMF finds Connections in Complex Data

Clare Lee[1] , Des Higham[1], Keith Vass[2], Dan Crowther[2,3]

4th September 2010

[1] Department of Mathematics & Statistics,
University of Strathclyde.
[2] Translational Medicine Research Collaboration,
University of Dundee.
[3] Pfizer Inc.

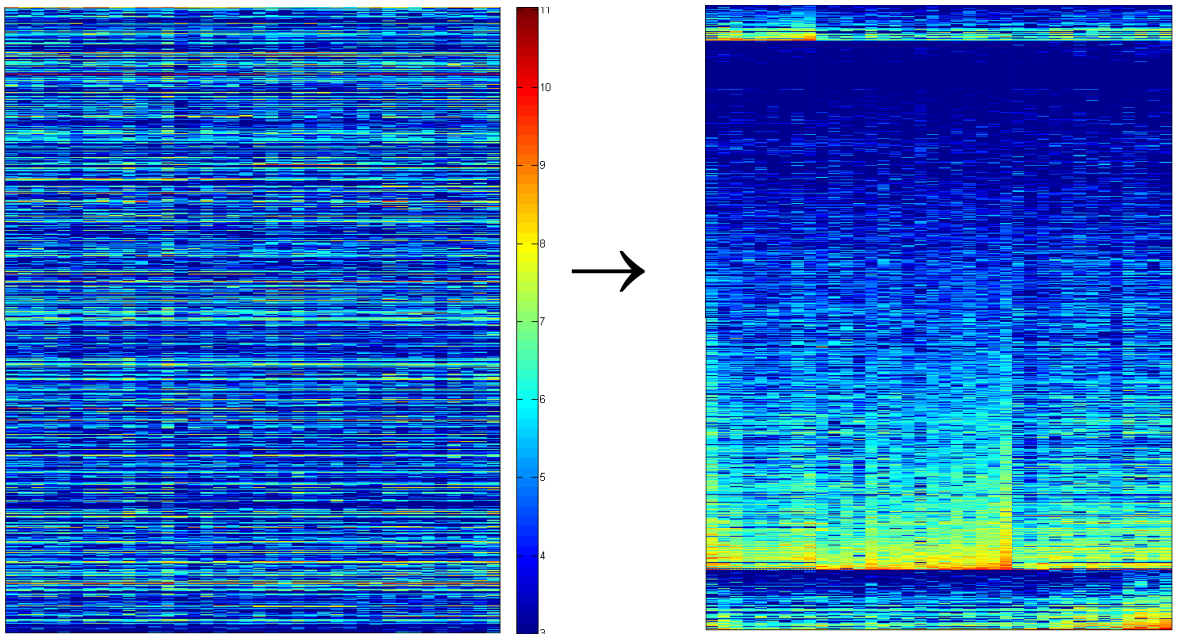


University of Strathclyde Glasgow

Contents

Aim

The aim is to take non-negative data sets, for example microarray data and to reorder or cluster the data to find hidden features using non-negative matrix factorisation (NMF).

$A =$ 

NMF Algorithms

There are many different algorithms to compute a NMF.

Typically they compute two factors so that

$$A = WH + error \text{ minimises } \|A - WH\|$$

with all entries of W and H being non-negative.
If A is of size $m \times n$, then W is $m \times k$ and H is $k \times n$, where $k \ll m$ or n.
In the iterative approach

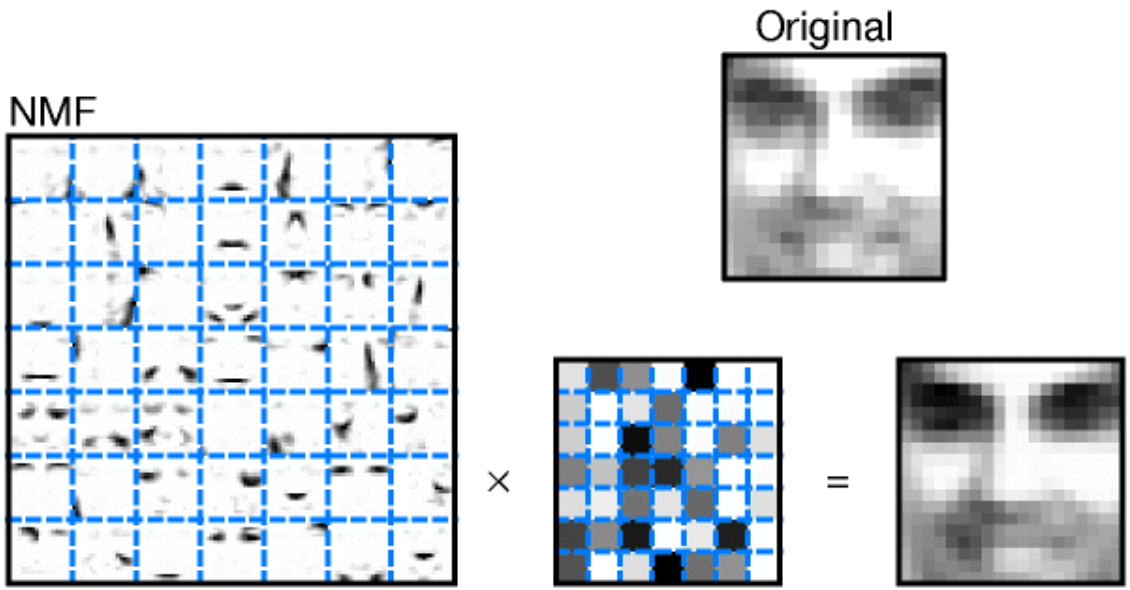
$$W_{i+1} = W_i \sum_{\text{samples}} \left(\begin{array}{c} \text{under/over-estimate} \\ \text{factor for} \\ \text{this sample} \end{array} \times \begin{array}{c} \text{importance} \\ \text{of sample} \\ \text{in cluster} \end{array} \right)$$

H is found in an analogous way.

Feature finding, Ordering and Clustering

$$A \approx WH = \sum_{j=1}^{k} w_j h_j^{\mathrm{T}},$$

for $W = [w_1, \ldots, w_k]$, and $H = [h_1, \ldots, h_k]^{\mathrm{T}}$. Each rank-one non-negative matrix $w_j h_j^{\mathrm{T}}$ expresses a "feature" of the data. As shown by Lee and Seung [Nature(1999)]



Feature finding, Ordering and Clustering

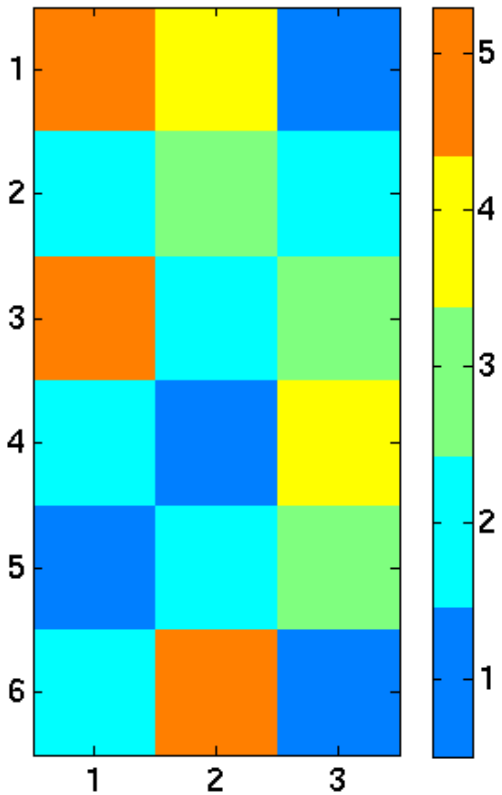
Therefore, for example in the microarray applications

- the columns of the first factor W are referred to as "eigen-genes"
- the rows of the second factor H are equivalently "eigen-samples"

Since each column/row expresses one feature we can locate this in the data by re-ordering the individual vectors to put the largest value in the bottom right corner.

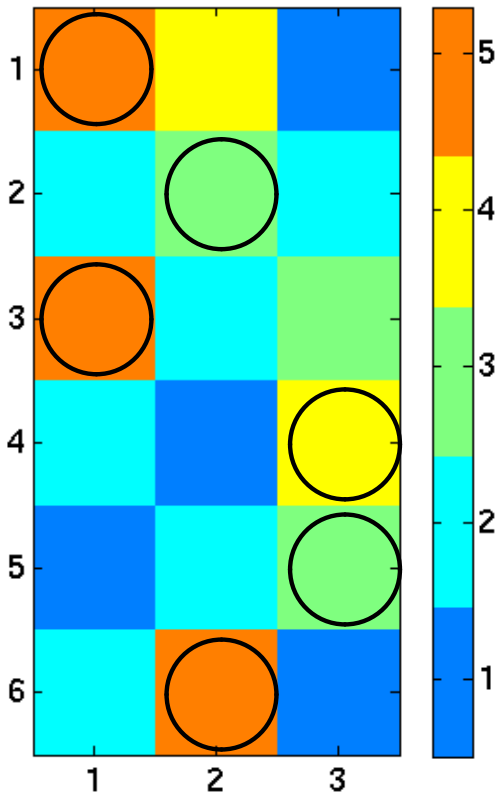
We can also combine these features into one ordering ...

Feature finding, Ordering and Clustering



Each row in W is assigned to a cluster corresponding to the largest element in that row.

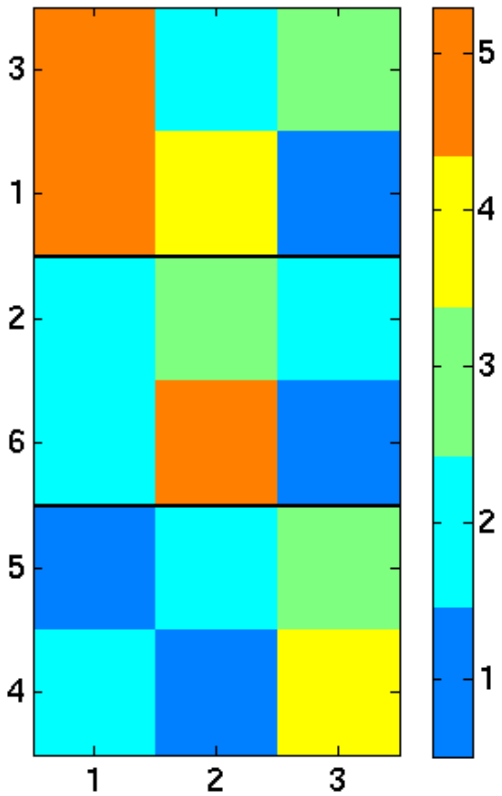
Feature finding, Ordering and Clustering



Each row in W is assigned to a cluster corresponding to the largest element in that row.

Row 1 is assigned to cluster 1
Row 2 is assigned to cluster 2
Row 3 is assigned to cluster 1
$\vdots$

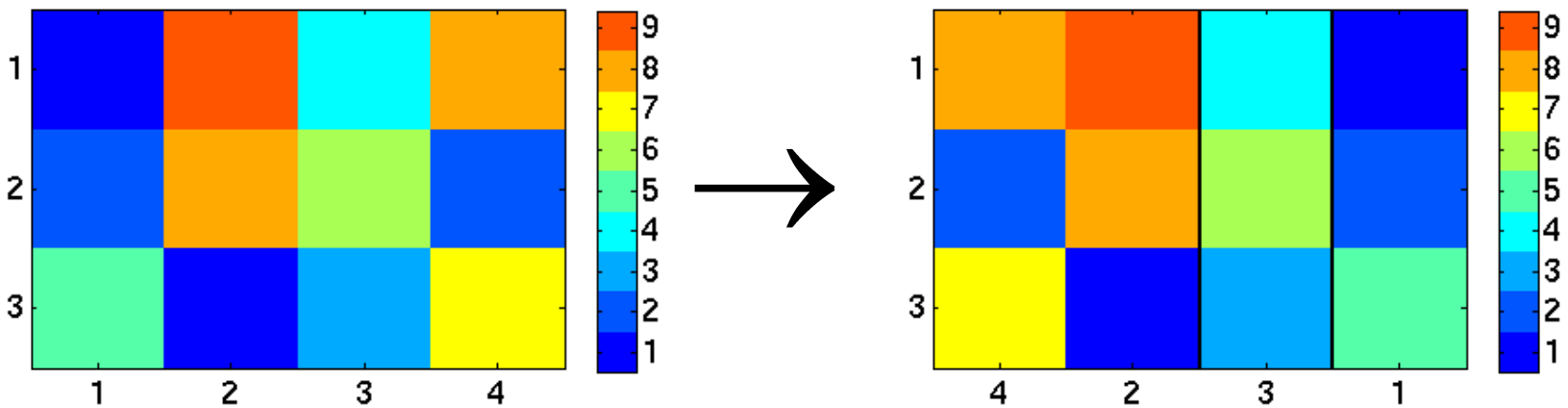
Feature finding, Ordering and Clustering



A row ordering then comes from stacking the clusters and sorting each cluster by size of that column.

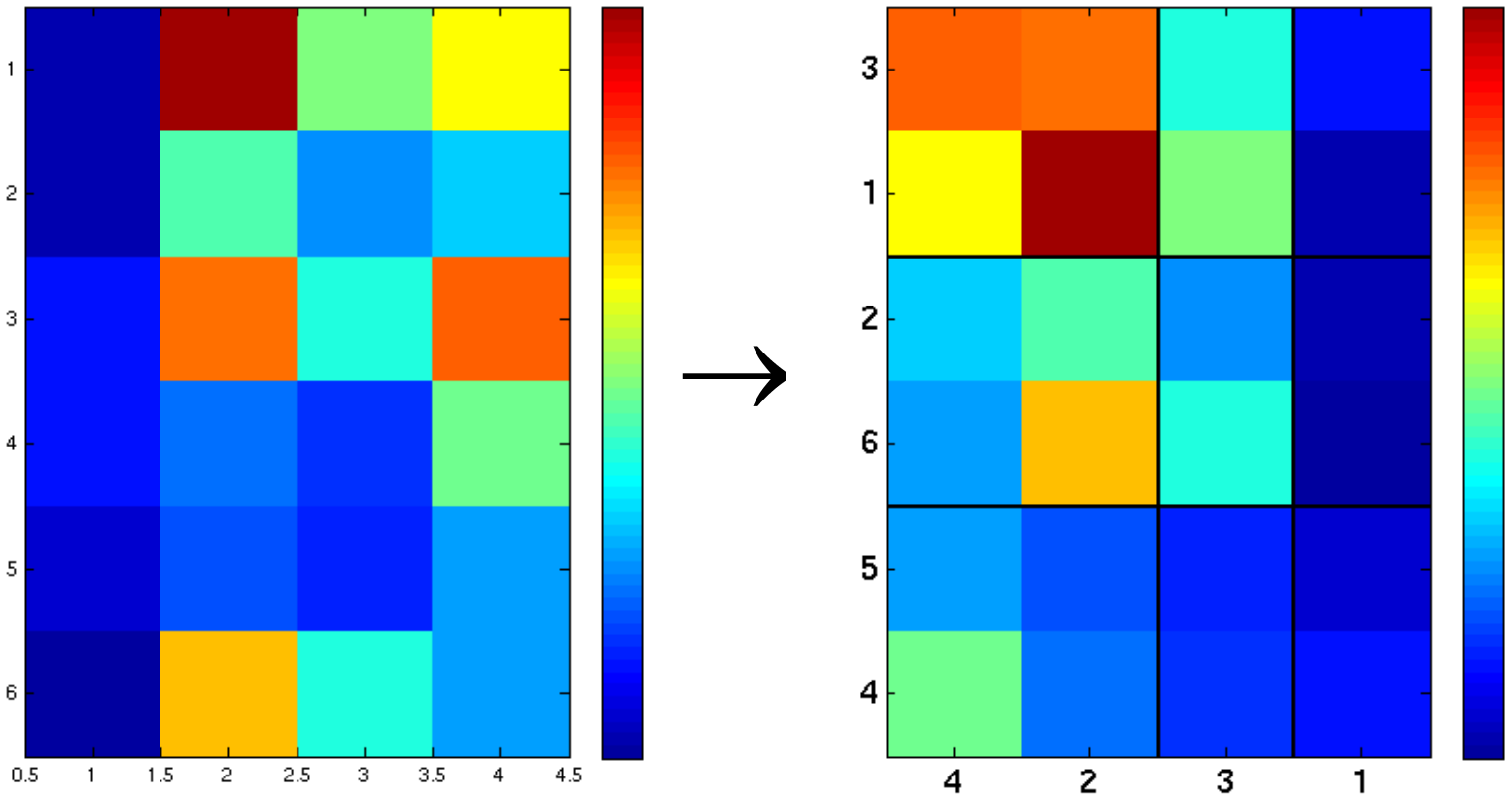
Feature finding, Ordering and Clustering

The same is done with H and the columns.

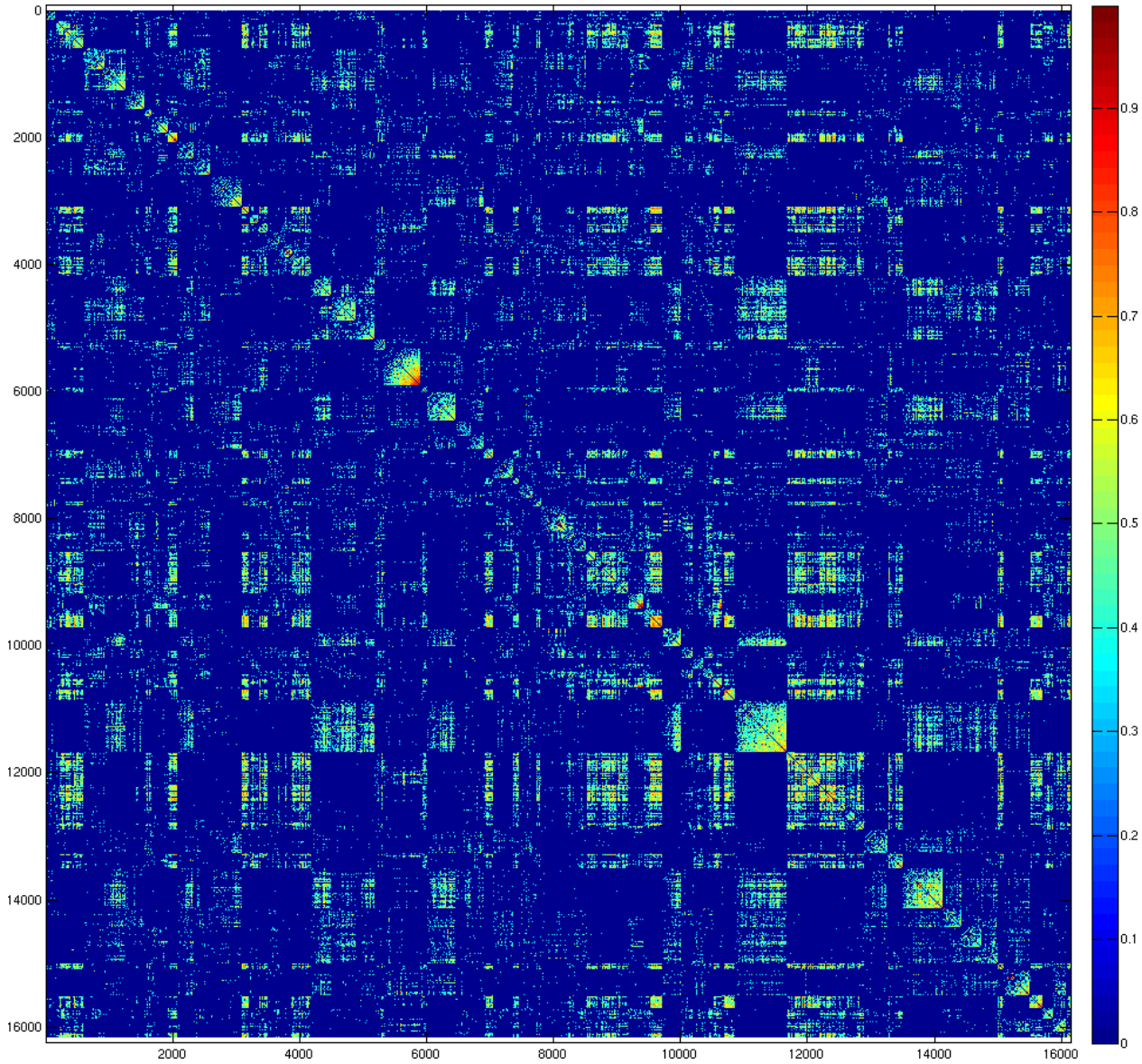


Feature finding, Ordering and Clustering

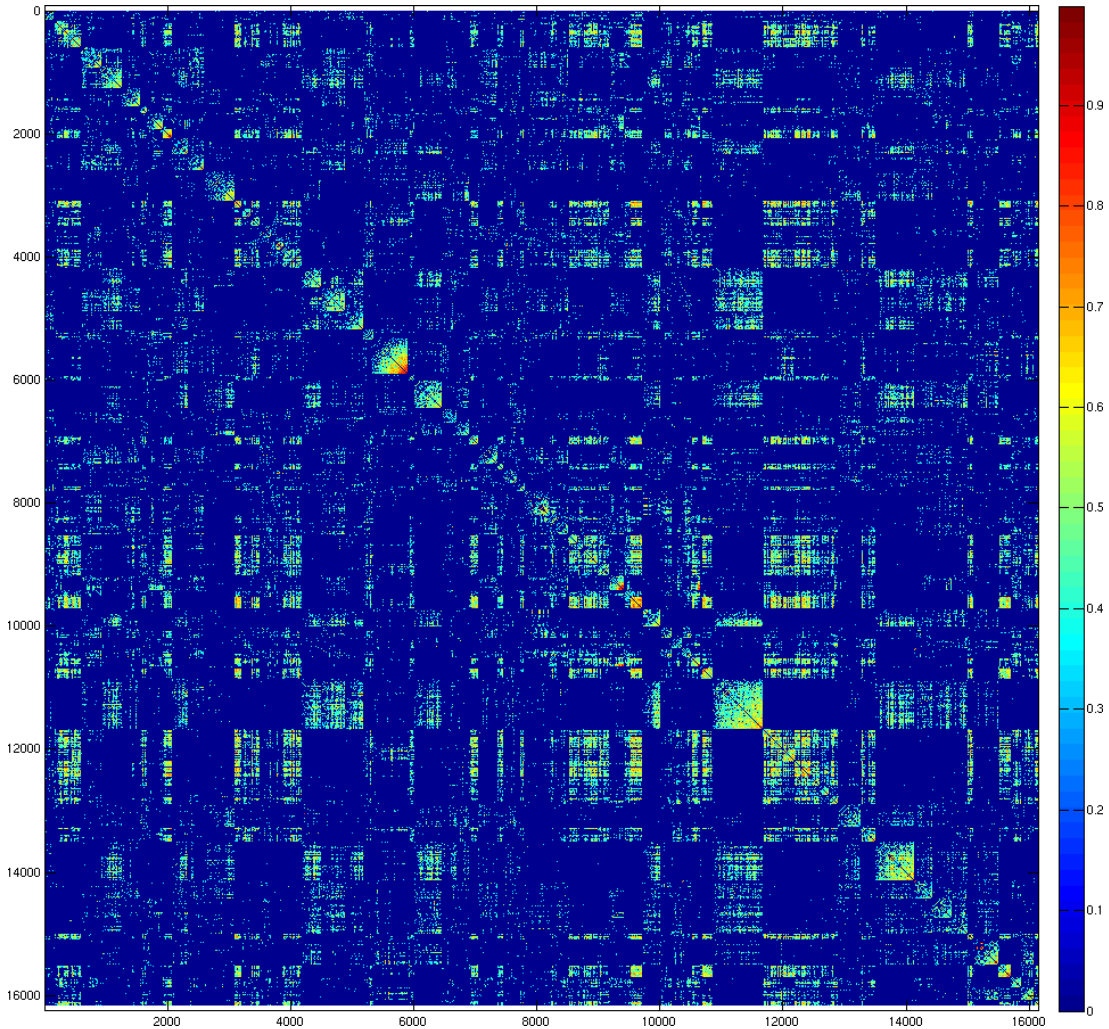
These orderings can then be applied to the original matrix.

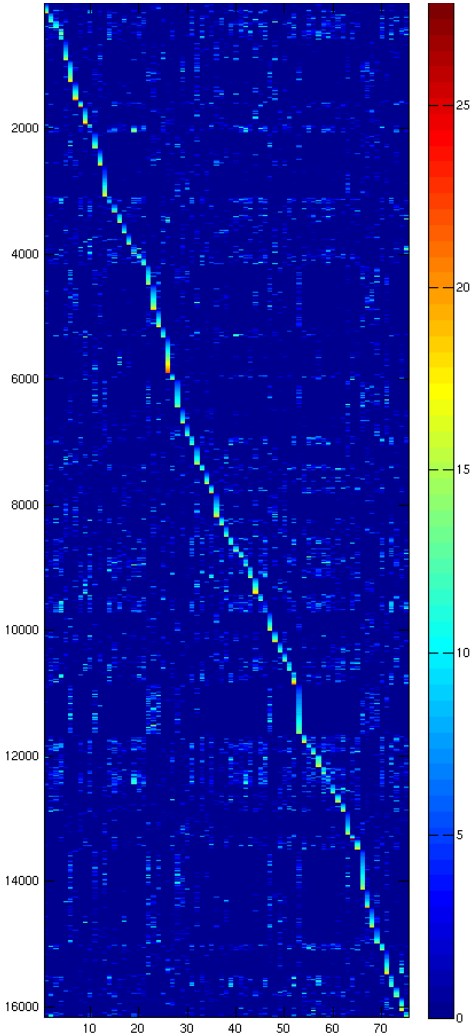


Colon Cancer Gene Correlation: k=75



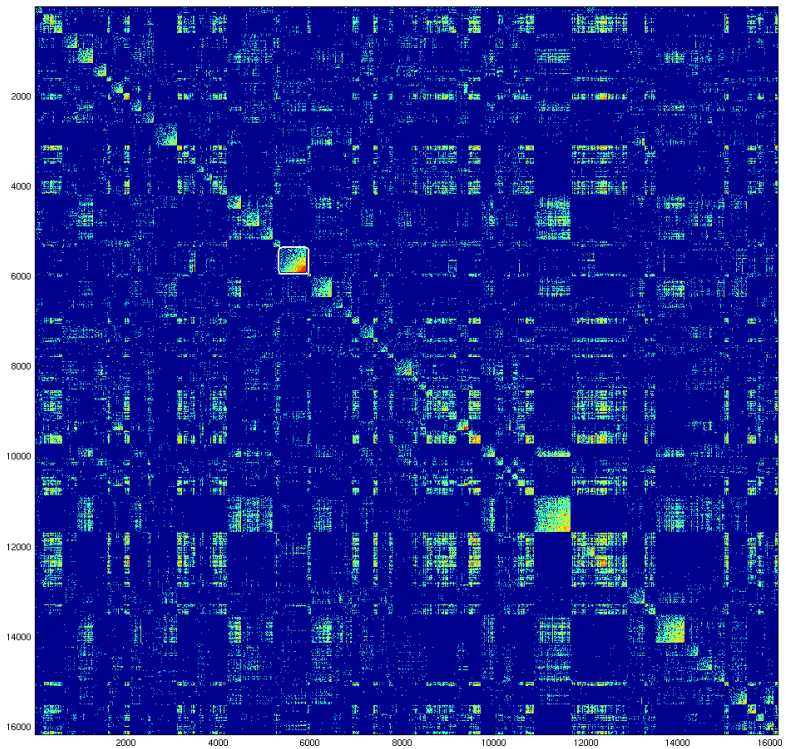
Colon Cancer Gene Correlation: k=75

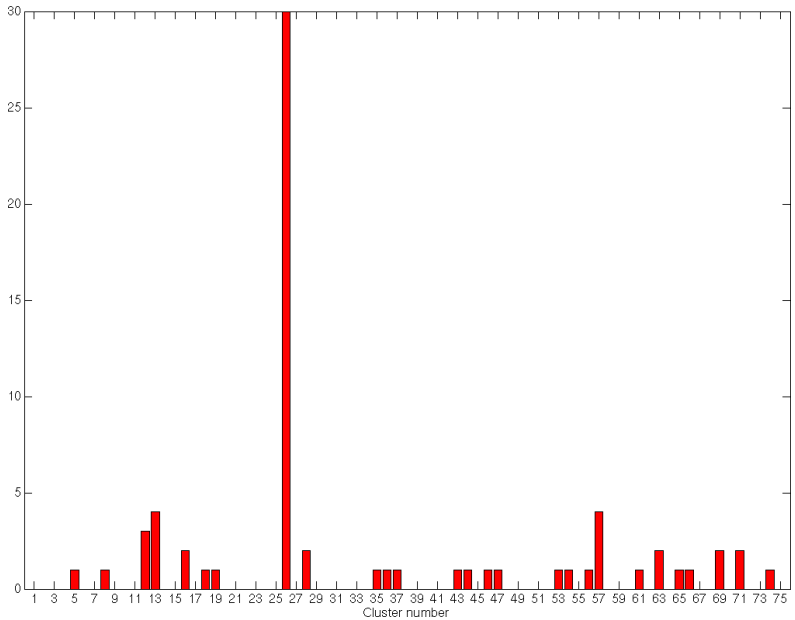




Comparing with known information

Genes suppressed by oncogene HRas



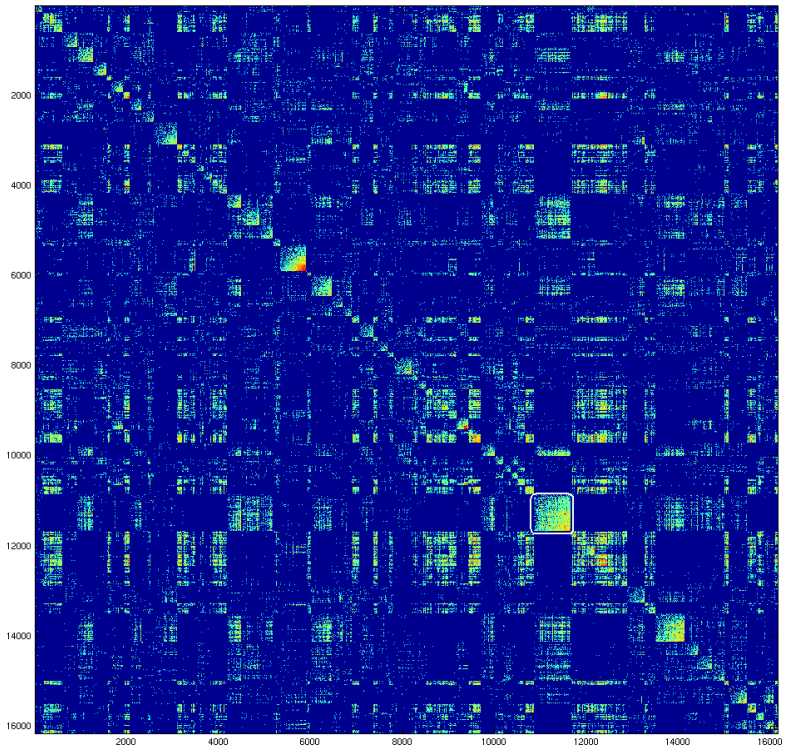


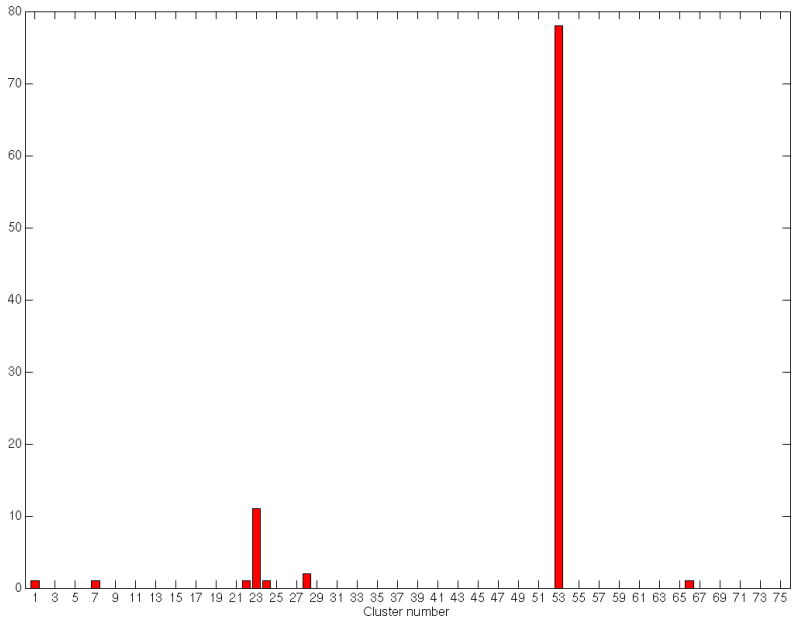
This cluster's "Ras signature" contains many proteins found in the "extracellular region". The cluster includes

ADAMTS5, C1S, CADM1, CH25H, COL11A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, CRISPLD2, DCN, EFEMP1, ELN, IGFBP3, LUM, MXRA5, POSTN, SPOCK1, SULF1

Comparing with known information

vec5 - probesets associated with cell division and DNA copying

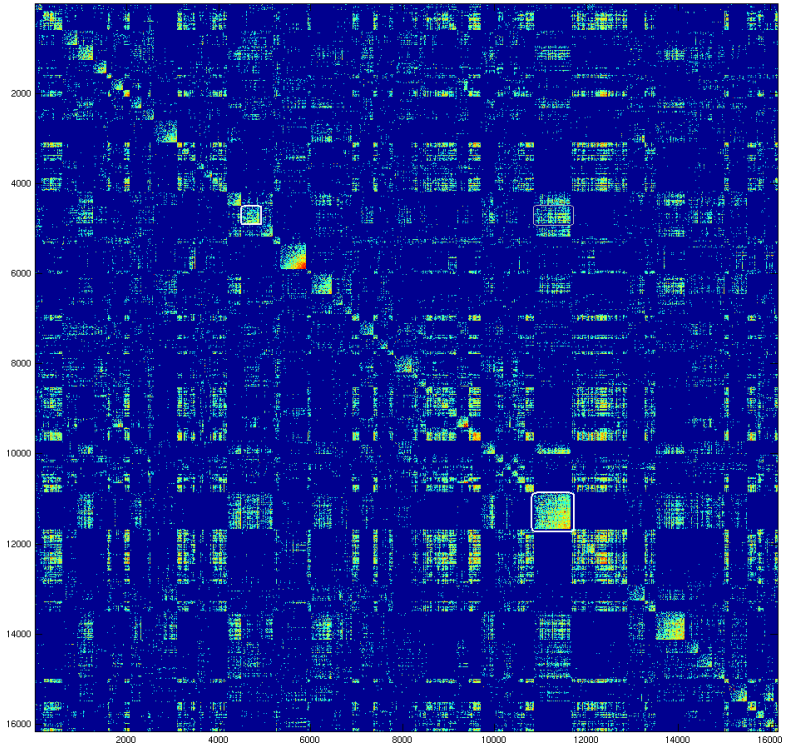


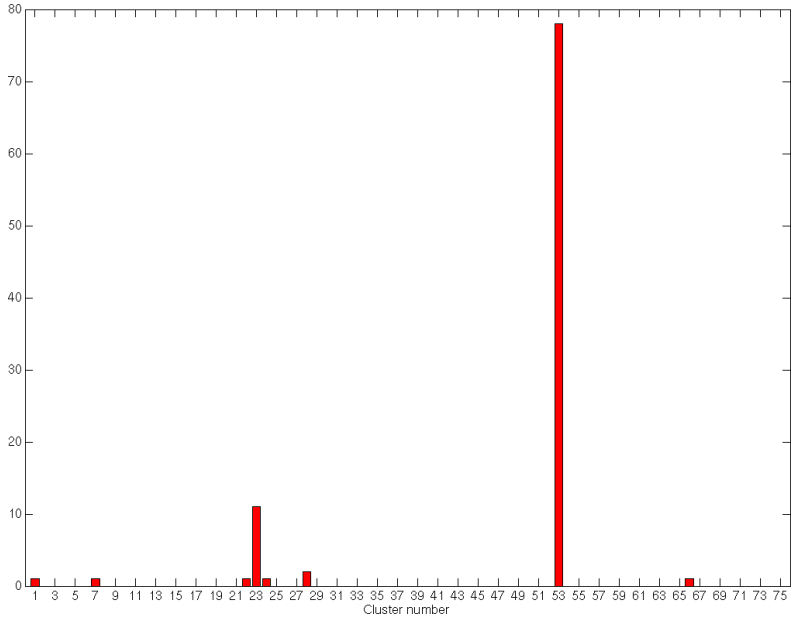


This cluster's "DNA replication and cell-division" set is enriched in proteins for the "nucleus" and the "mitochondrion".

Comparing with known information

vec5 - probesets associated with cell division and DNA copying

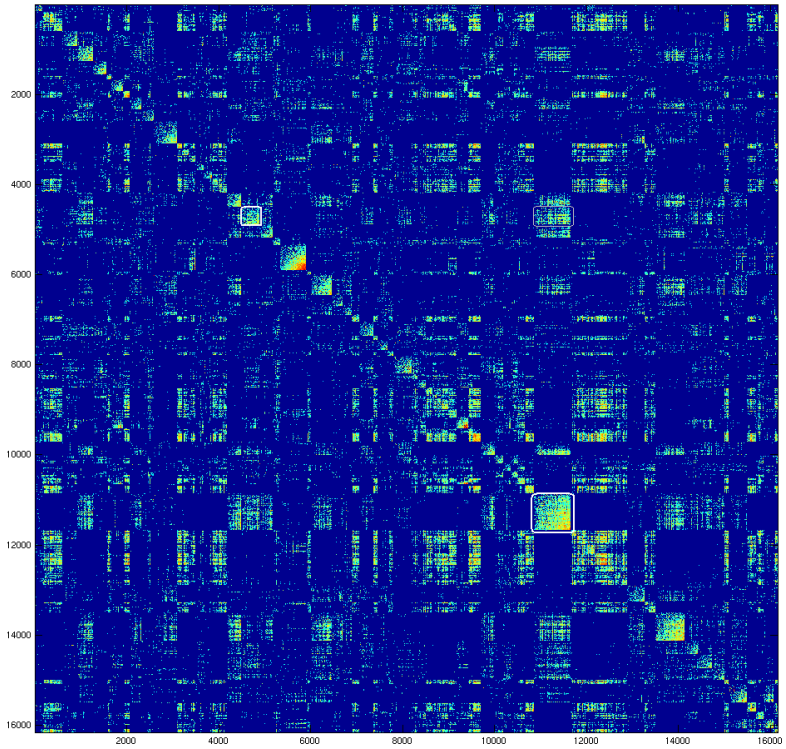


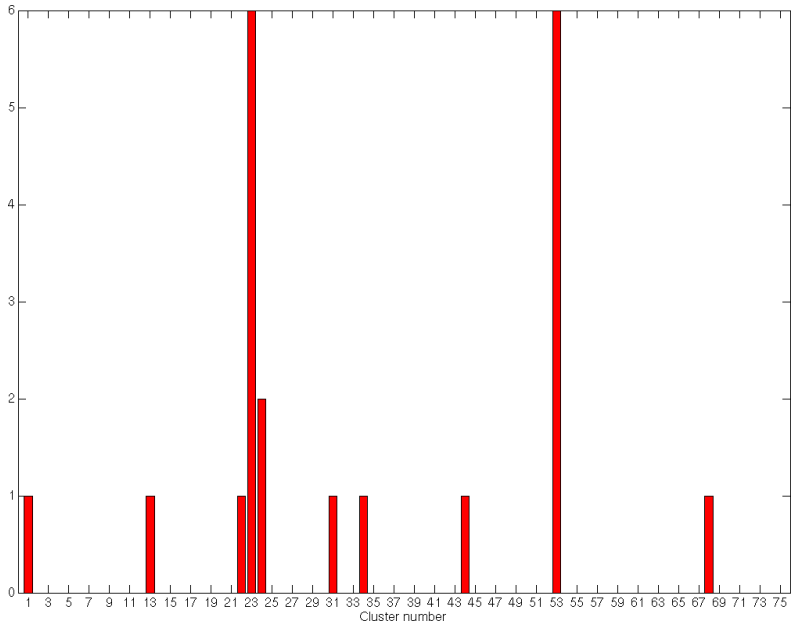


This cluster's "DNA replication and cell-division" set is enriched in proteins for the "nucleus" and the "mitochondrion".

Comparing with known information

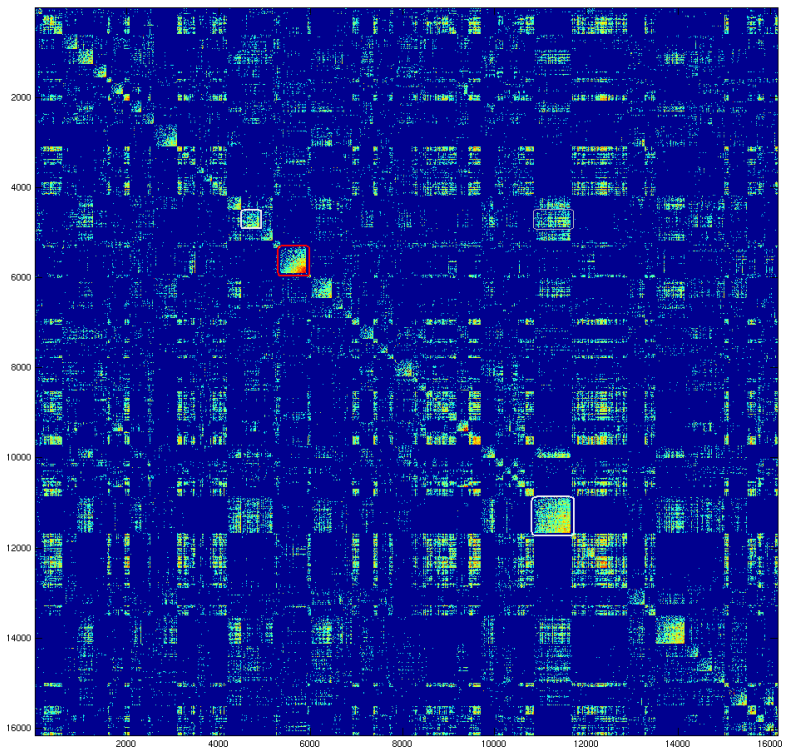
C2 set of genes associate with Notch pathway being active

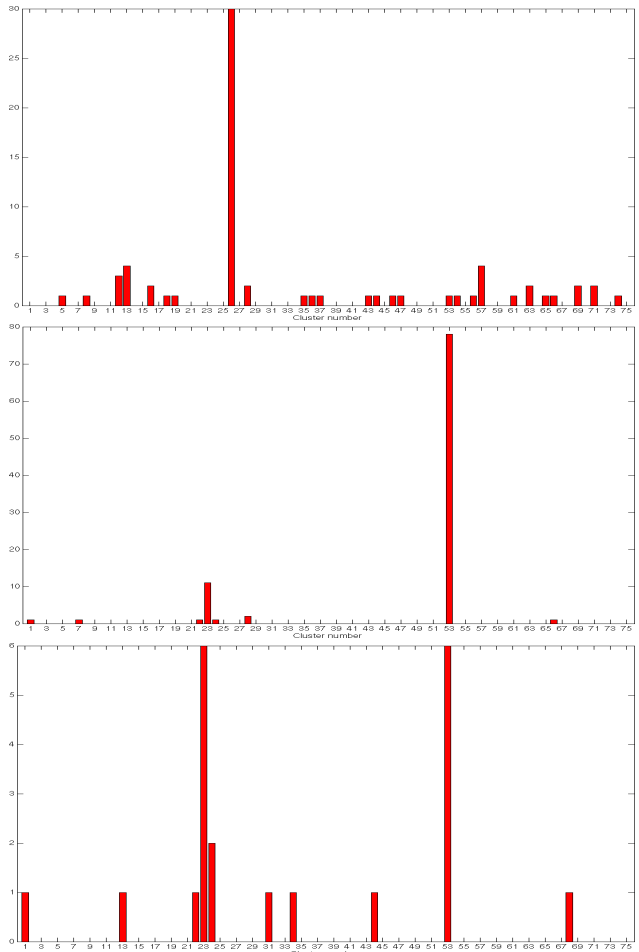




Comparing with known information

HRas suppressed, vec5 and C2 genes





Using Multiple Data Sources

In some situations it may be advantageous to use more than one data source to improve our results or to look for differences and similarities between data sources.

For this there is Simultaneous NMF to factorise data matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times n}$ so that

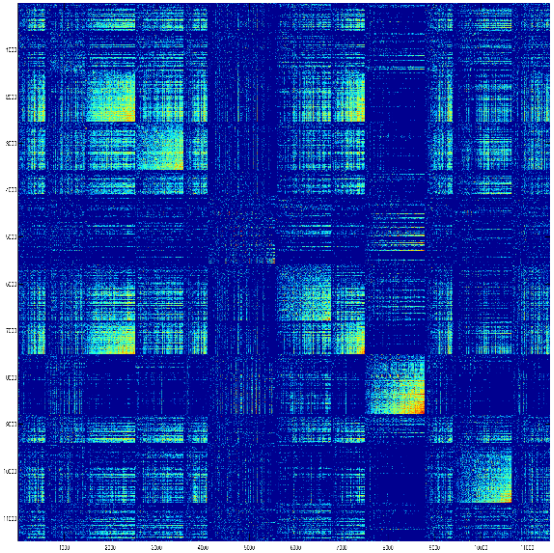
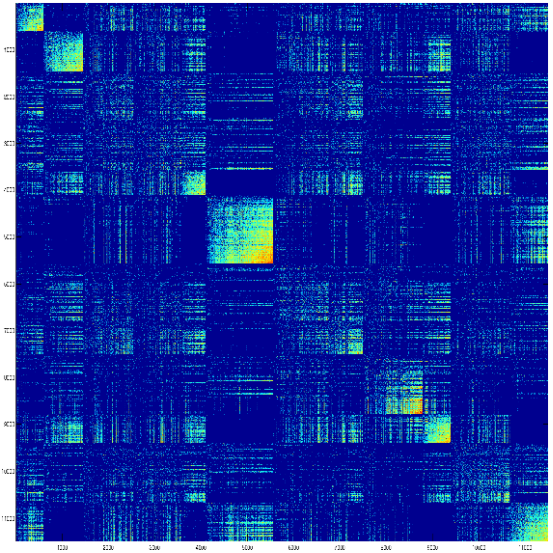
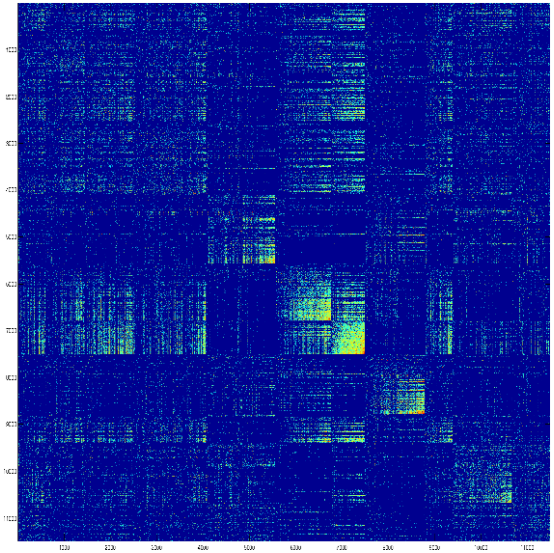
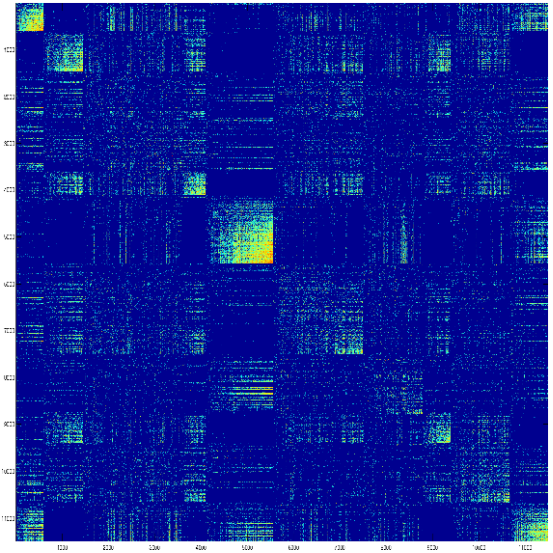
$$A \approx WH \text{ and } B \approx SH$$

with $W \in \mathbb{R}^{m \times k}, S \in \mathbb{R}^{p \times k}$ and $H \in \mathbb{R}^{k \times n}$. Producing a matching ordering/clustering of the columns of the two matrices.

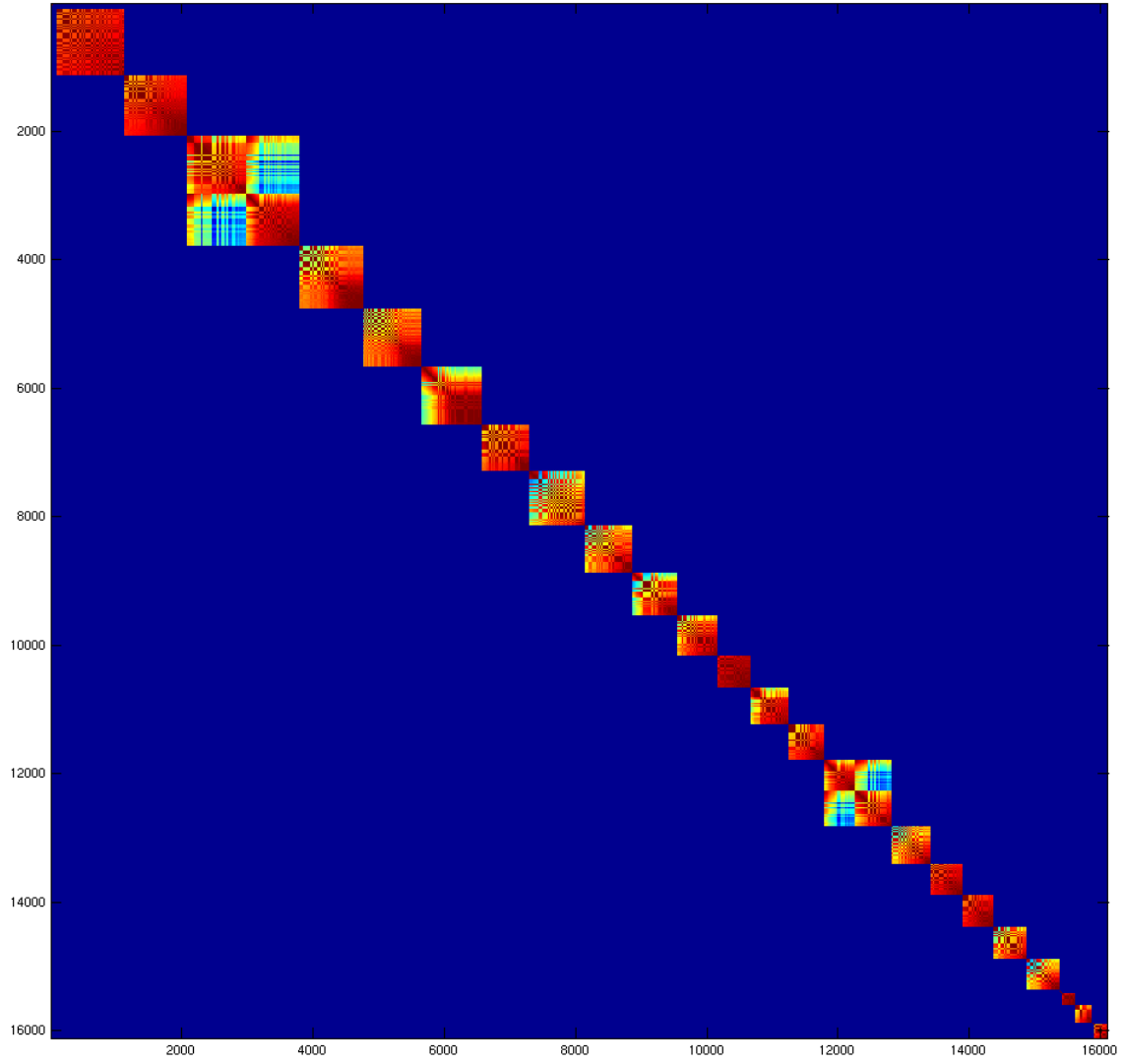
[Badea, *Proc.Pacific Symp.BioInf.*(2008))]

We have extended this further to take any number of matrices

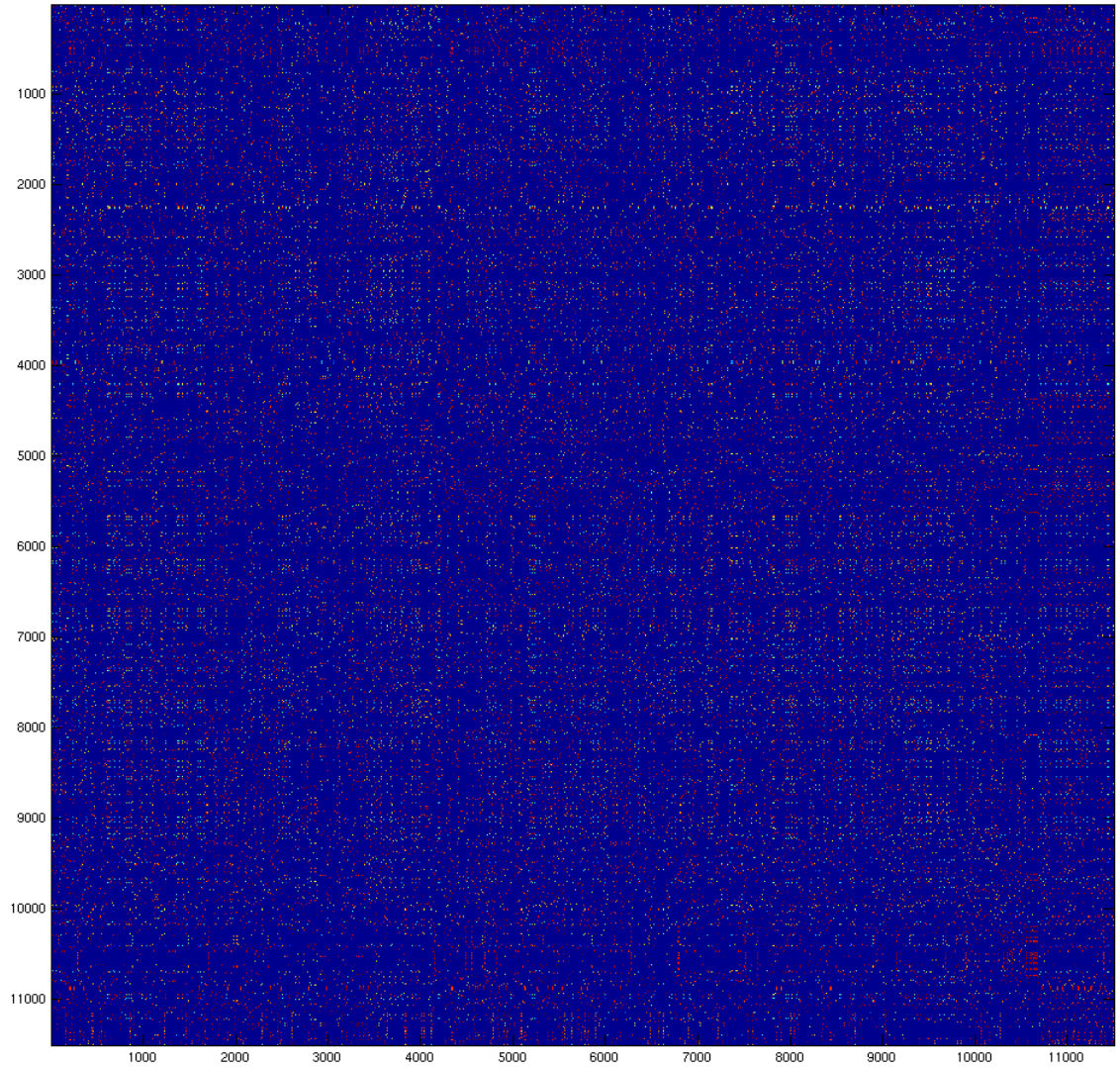
Colon Cancer Correlation Matrices: Four data sets $k = 12$

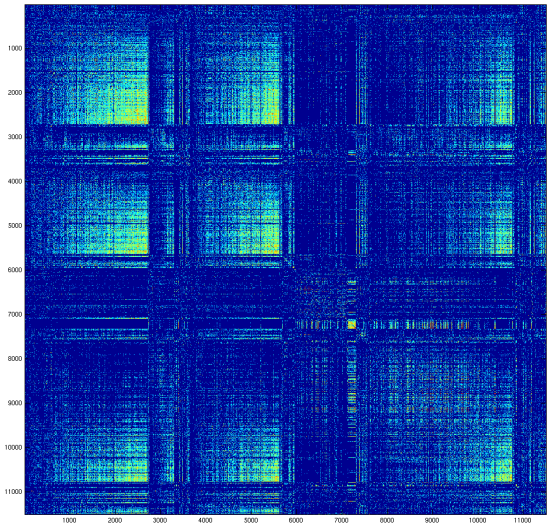
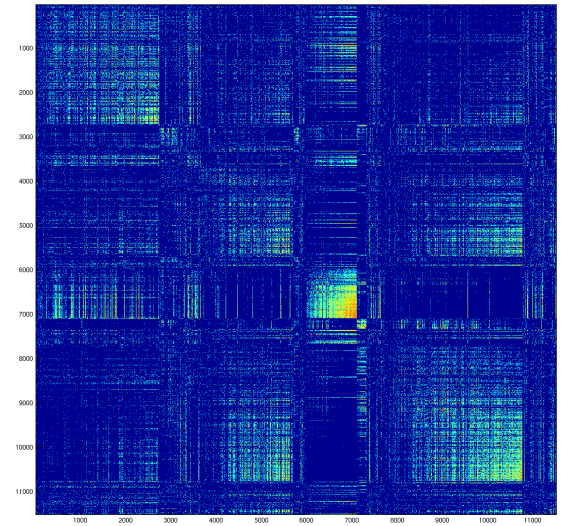
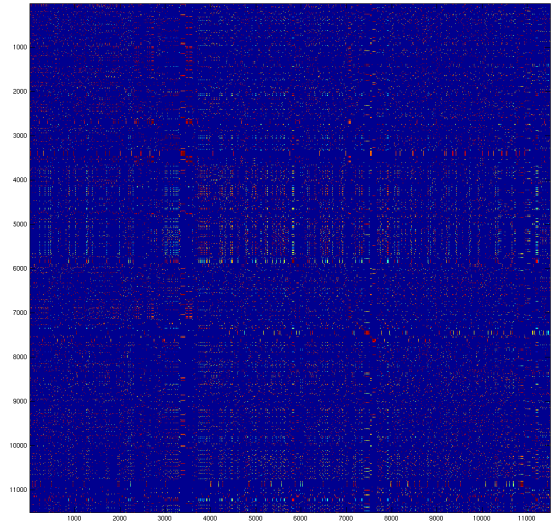
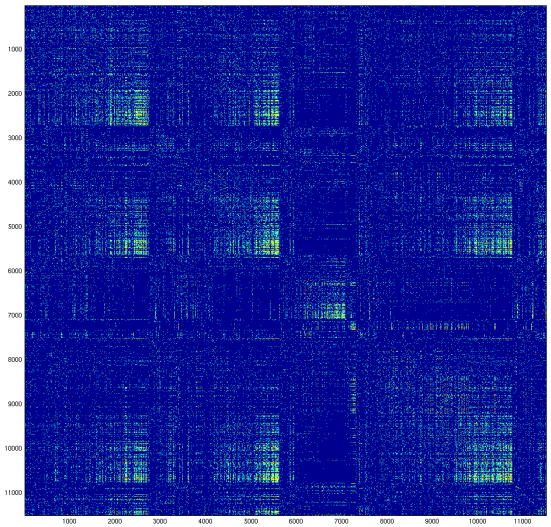
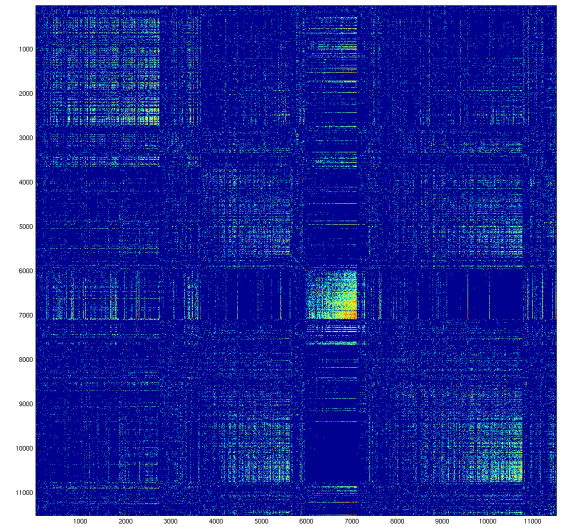
Chromosomal Location



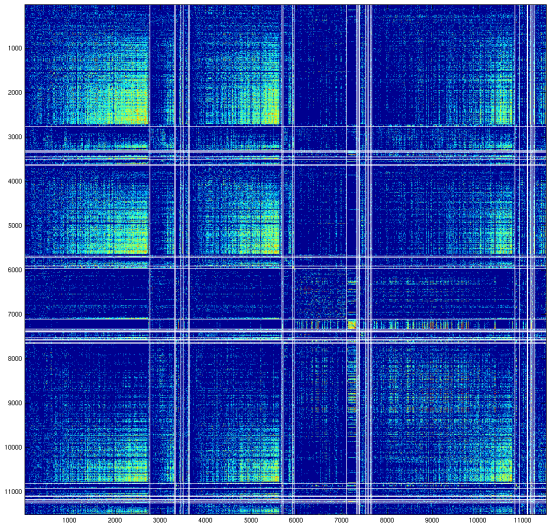
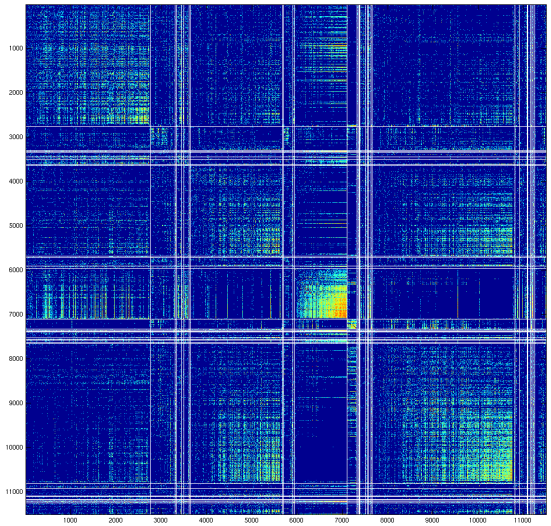
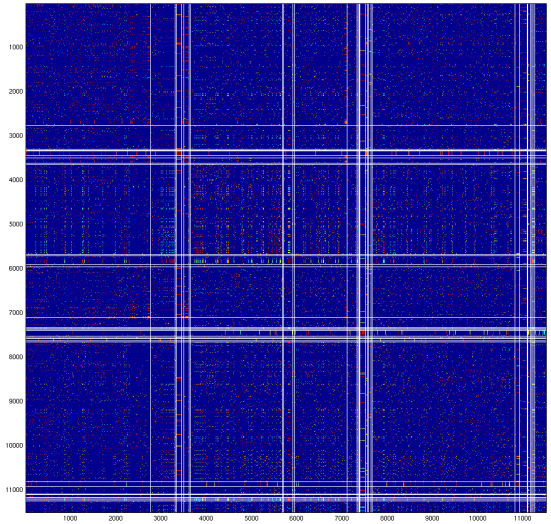
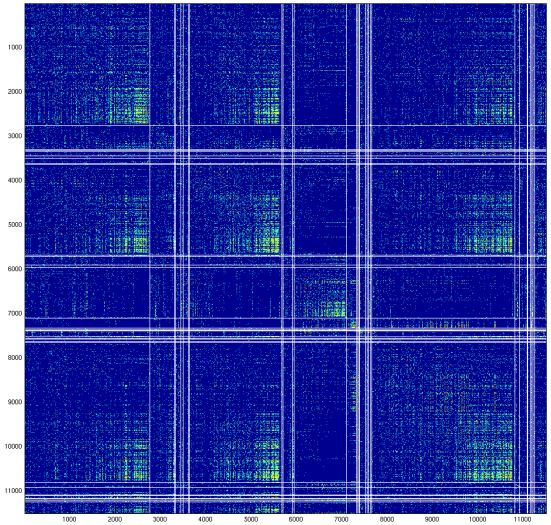
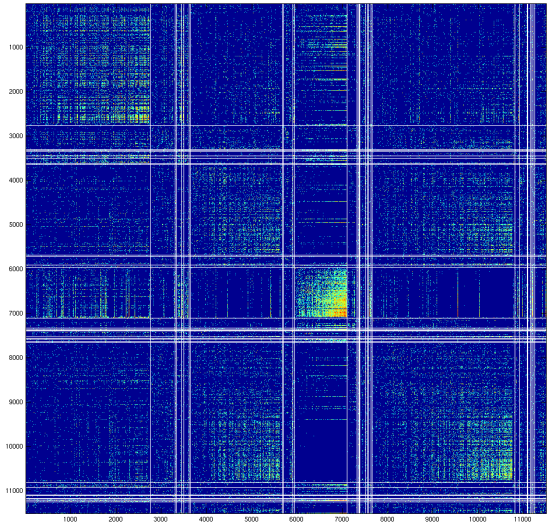
Colon Cancer Correlation Matrices: Four data sets $k = 12$



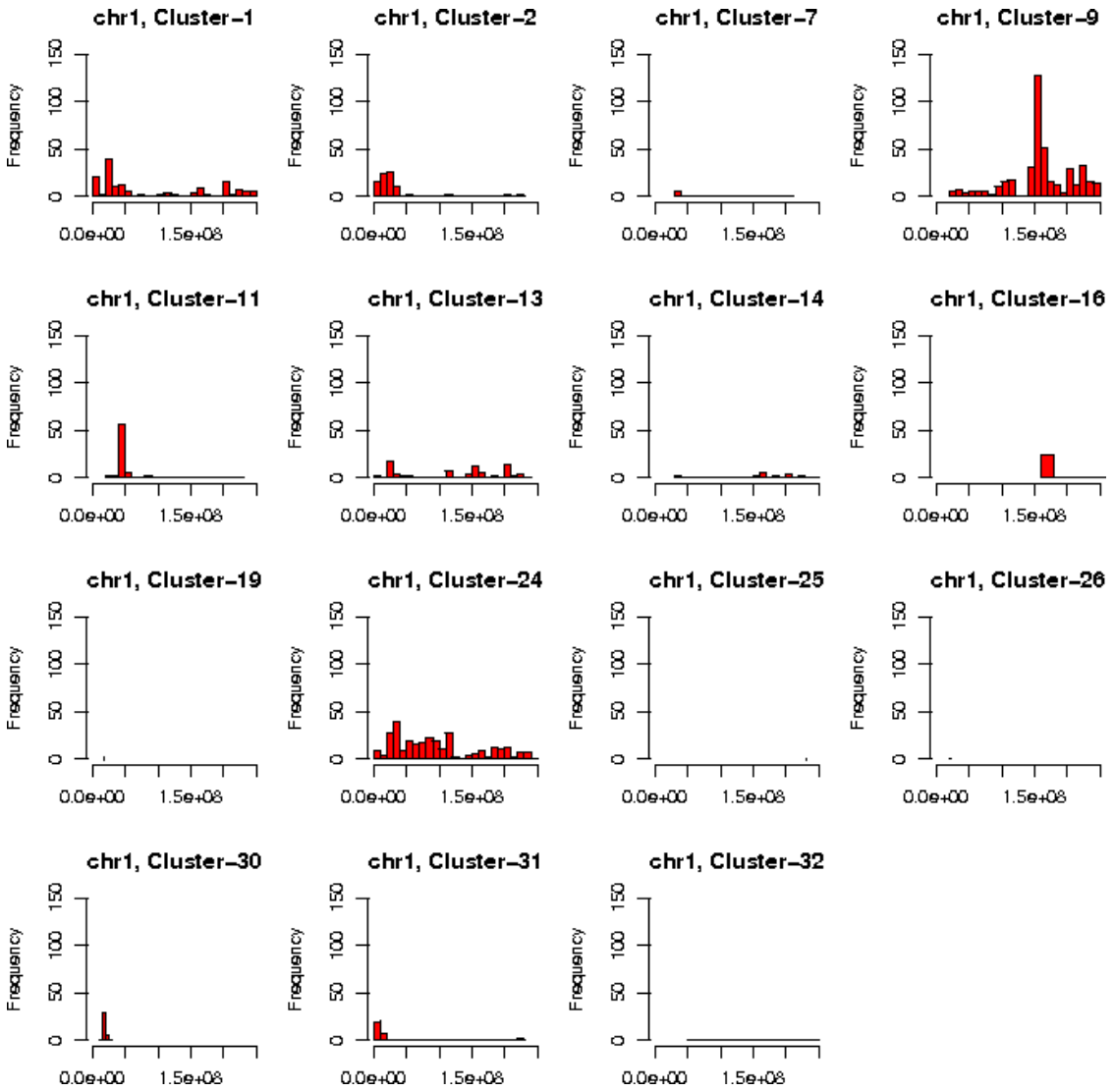
Colon Cancer: Adding extra information, $k = 32$

Colon Cancer: Adding extra information, $k = 32$

Colon Cancer: Adding extra information, $k = 32$



Looking to the literature

Chromosome 1 p34.1-p34.2 – cluster 11

Molecular Biology, Vol. 34, No. 3, 2000, pp. 337–344. Translated from Molekulyarnaya Biologiya, Vol. 34, No. 3, 2000, pp. 397–405.

GENOMICS. PROTEOMICS. BIOINFORMATICS

UDC 575:599.9

Deletion of the *Alu-VpA/MycL1* ~~(1p34.3)~~ (1p34.2) Locus is a Negative Prognostic Sign in Human Colorectal Cancer

K. N. Kashkin, A. G. Perevoschikov, A. V. Nikolaev, D. A. Turbin, and E. W. Fleischman

Blokhin Cancer Research Center, Russian Academy of Medical Sciences, Moscow, 115478 Russia;
E-mail: kirill@online.ru

Received February 3, 2000

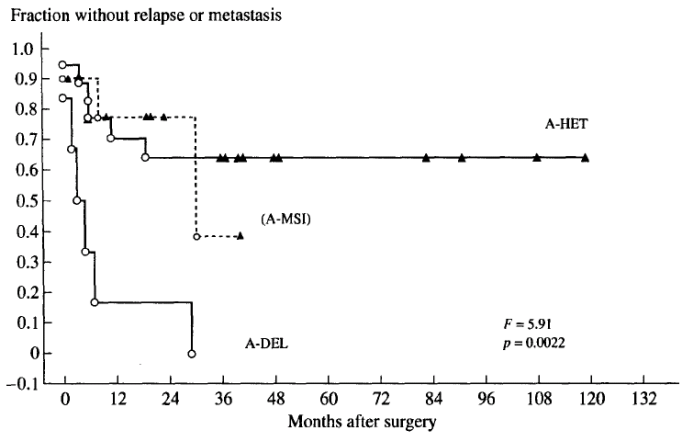
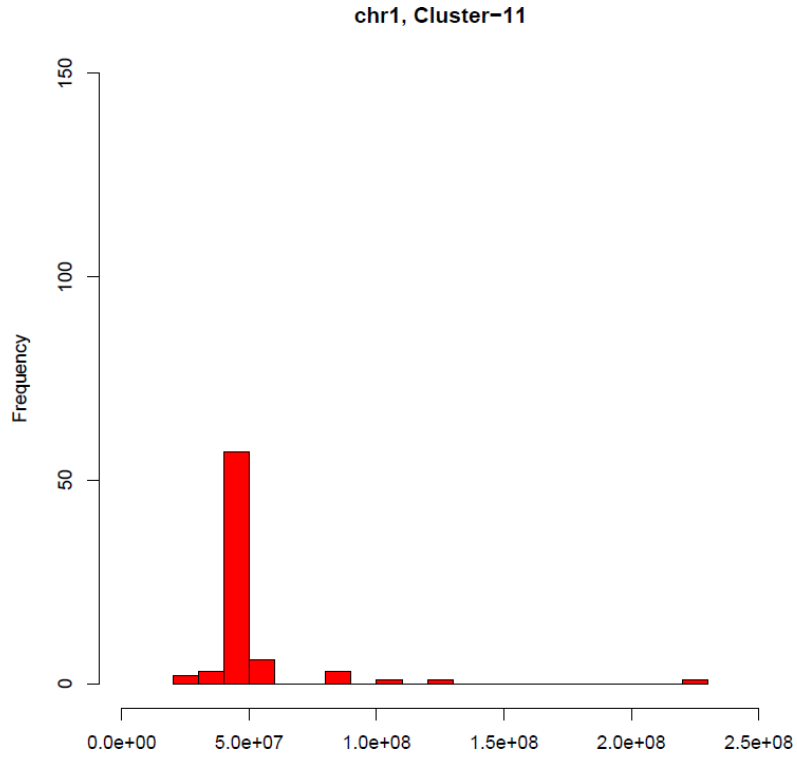


Fig. 4. Relapse-free life span of patients with *Alu-VpA/MycL1* disturbances in the Kaplan–Meier assessment. (A-DEL) Loss of the *Alu-VpA/MycL1* allele; (A-MSI) *Alu-VpA/MycL1* instability; (A-HET) heterozygous tumors indistinguishable in *Alu-VpA/MycL1* from the normal mucosa. (○) Relapse or metastases; (▲) censored without relapse; (*F*) Cox test for A-HET and A-DEL (A-MSI was not assessed because of the short terms of observation).



Looking to the literature

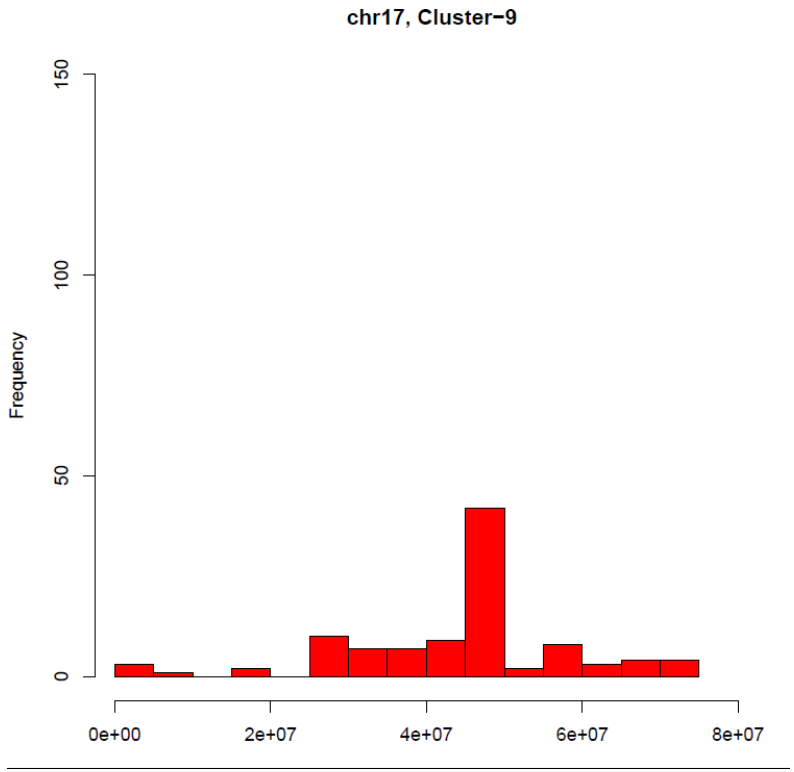
Chromosome 17 q21.2-q23.2 – cluster 9

Int. J. Cancer (Pred. Oncol.): 89, 1–7 (2000)
Publication of the International Union Against Cancer

GENOMIC ALTERATIONS (LOH, MI) ON CHROMOSOME 17q21–23 AND PROGNOSIS OF SPORADIC COLORECTAL CANCER

Christophe R. BERNEY[1,3]*, Richard J. FISHER[2], Jia-lin YANG[1,3], Pamela J. RUSSELL[3] and Philip J. CROWE[1]
[1]*Department of Surgery, Prince of Wales Hospital, University of New South Wales, Randwick, New South Wales, Australia*
[2]*Department of Radiation Oncology, Prince of Wales Hospital, University of New South Wales, Randwick, New South Wales, Australia*
[3]*Oncology Research Centre, Prince of Wales Hospital, University of New South Wales, Randwick, New South Wales, Australia*

In conclusion, fluorescent PCR technology coupled with an automated DNA sequencer appeared to be a very accurate and reliable method for detection of microsatellite alterations in genomic DNA extracted from paraffin-embedded material. Genomic alterations in the 17q21–23 region may affect prognosis of CRC as well as regulation of the nm23 protein expression *via* an unknown underlying mechanism. Finally, the area flanking the *D17S579* and *MPO* loci is likely to contain potential tumour suppressor gene(s) in which mutational inactivation play(s) a significant role for development and/or progression of at least some sporadic colorectal tumours.



Looking to the literature

Chromosome 17 q25 – cluster 12

Gastroenterology
Volume 114, Issue 6, June 1998, Pages 1206-1210

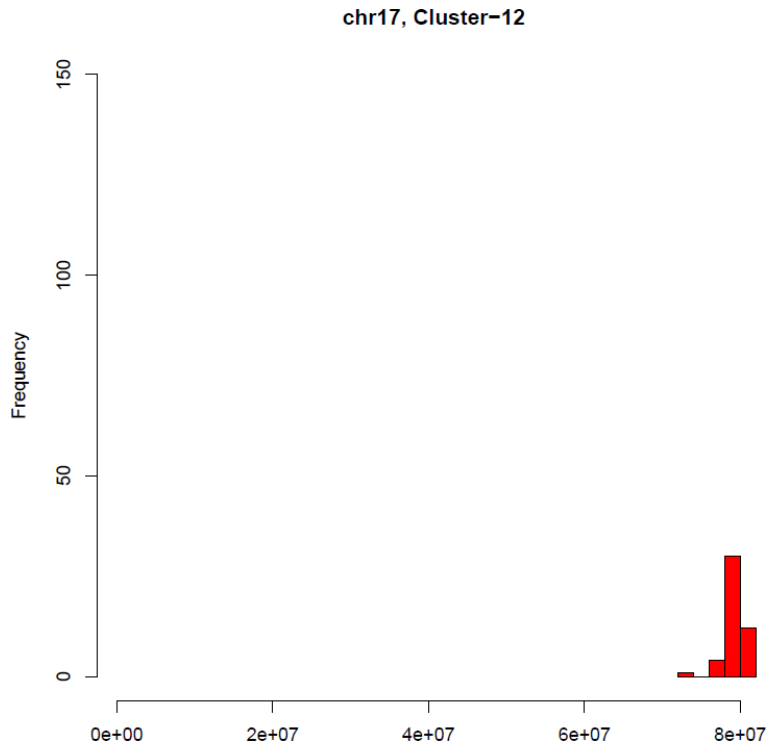
AGA INSTITUTE
AGA Education, Practice and Research

doi:10.1016/S0016-5085(98)70426-3 | How to Cite or Link Using DOI

Alimentary Tract

Tylosis esophageal cancer locus on chromosome 17q25.1 is commonly deleted in sporadic human esophageal cancer☆, , ☆☆

Takeshi Iwaya*, Chihaya Maesawa*, Satoshi Ogasawara‡ and Gen Tamura*
Departments of *Pathology and ‡Surgery, Iwate Medical University School of Medicine, Uchimaru, Morioka, Japan
Received 29 October 1997; accepted 17 February 1998. Available online 5 November 2005.



Looking to the literature

Chromosome 11 q13 – cluster 5

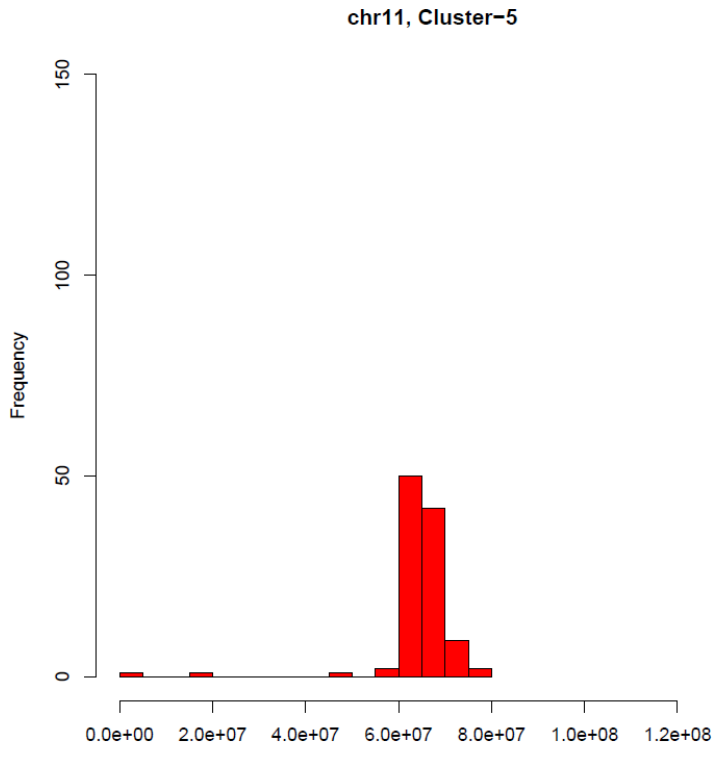
GENES, CHROMOSOMES & CANCER 22:130–137 (1998)

Deletion Mapping of Endocrine Tumors Localizes a Second Tumor Suppressor Gene on Chromosome Band 11q13

Rita Chakrabarti,[1] Eri S. Srivatsan,[1]* Thomas F. Wood,[1] Patricia J. Eubanks,[2] Sam A. Ebrahimi,[1] Richard A. Gatti,[3] Edward Passaro, Jr.,[1] and Mark P. Sawicki[1]

[1]Department of Surgery, VAMC West Los Angeles, UCLA School of Medicine, Los Angeles, California
[2]Department of Surgery, Harbor/UCLA Medical Center, Torrance, California
[3]Department of Pathology, UCLA School of Medicine, Los Angeles, California

Multiple endocrine neoplasia type 1 syndrome (*MEN1*, MIM 131100), an autosomal dominant disease, is characterized by parathyroid hyperplasia, pancreatic endocrine tumors, and pituitary adenomas. These tumors also occur sporadically. Both the familial (*MEN1*) and the sporadic tumors reveal loss of heterozygosity (LOH) for chromosome band 11q13 sequences. Based on prior linkage and LOH analyses, the *MEN1* gene was localized between *PYGM* and D11S460. Recently, the *MEN1* gene (menin) has been cloned from sequences 30-kb distal to *PYGM*. We performed deletion mapping on 25 endocrine tumors (5 *MEN1* and 20 sporadic) by using 21 polymorphic markers on chromosome band 11q13. Of these, two (137C7A, 137C7B) were derived from *PYGM*-containing BAC (bacterial artificial chromosome-137C7) sequences, one from *INT2*-containing cosmid sequences and the marker D11S4748, a $(CA)_{20}$ repeat marker that was developed by us. The LOH analysis shows that the markers close to the *MEN1* (menin) gene were not deleted in three of the tumors. These tumors, however, showed LOH for distal markers. Thus, the data suggest the existence of a second tumor suppressor gene on chromosome band 11q13. *Genes Chromosomes Cancer 22:130–137, 1998.*



Both MEN1 and HRASLS3, known tumour suppressors are included in the identified cluster

Discussion

Linking chromosomal information with correlation analysis we find that;

- the chromosomal information changes the gene-expression only clustering.
- the gene-expression data only links *some* of the gene neighbourhoods.
- many of the clusters have been previously described in either colorectal or another form of cancer.

Ongoing work

The ongoing investigations include

- Looking at the CRC gene-expression directly rather than the correlation matrices.
- Considering how normalising the data could affect the ordering.
- Considering ways of picking an "optimal" number of clusters if one exists