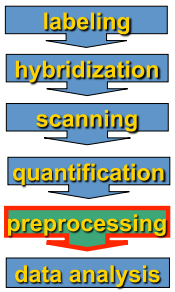
# Dimension reduction of gene expression profiles

Vincent Detours
vdetours@ulb.ac.be

• IRIBHM, UNIVERSITE LIBRE DE BRUXELLES

## Hierarchical clustering

## The expression matrix

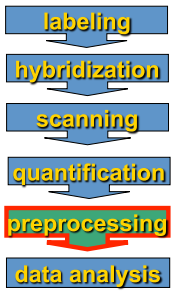



The result of all this is a matrix of expression ratios (tumor/healthy) for M genes x N tumors (e.g. 25000 x 100)

N tumors

| M genes | T1 | T2 | … | TN |
|---|---|---|---|---|
| G1 | 1.5 | -1.0 | … | 0.3 |
| G2 | -0.5 | 3.0 | … | 0.2 |
| G3 | 0.0 | 0.0 | … | -0.1 |
| … | … | … | … | … |
| GM | 2.6 | -0.2 | … | 1.1 |

## The expression matrix

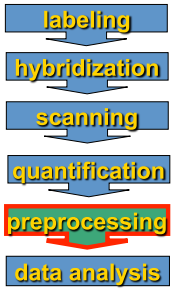



Look at the N tumors as points in the M-dimensional 'gene space'

N tumors

| M genes | T1 | T2 | … | TN |
|---|---|---|---|---|
| G1 | 1.5 | -1.0 | … | 0.3 |
| G2 | -0.5 | 3.0 | … | 0.2 |
| G3 | 0.0 | 0.0 | … | -0.1 |
| … | … | … | … | … |
| GM | 2.6 | -0.2 | … | 1.1 |

## The expression matrix



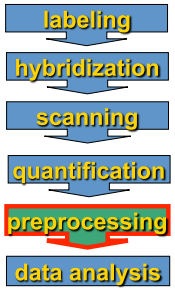

Look at the M genes as points in the N-dimensional 'phenotype space'

N tumors

M genes

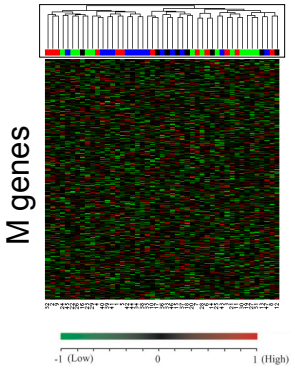
| | T1 | T2 | … | TN |
|---|---|---|---|---|
| G1 | 1.5 | -1.0 | … | 0.3 |
| G2 | -0.5 | 3.0 | … | 0.2 |
| G3 | 0.0 | 0.0 | … | -0.1 |
| … | … | … | … | … |
| GM | 2.6 | -0.2 | … | 1.1 |

## The expression matrix

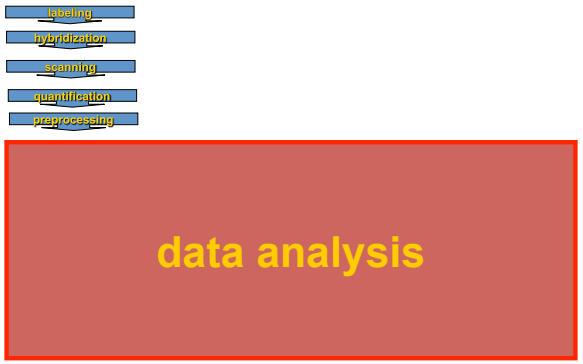



N tumors

M genes




## Biological investigation starts with data analysis

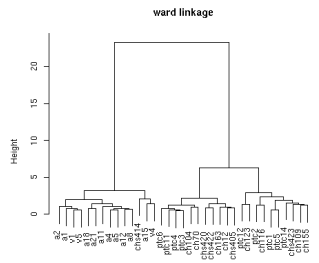



## Unsupervised and supervised machine learning

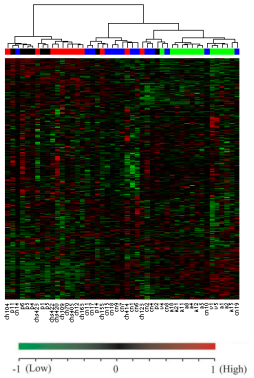
- With unsupervised learning the machine *discovers* a class structure in the data that is *not* known a priori
- With supervised learning the class structure of the data is *given* beforehand, and the machine finds gene expression patterns that *classify* tumors according to this structure

8

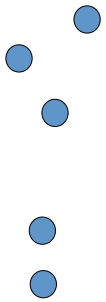
## Hierarchical clustering uncovers sub-groups in the data



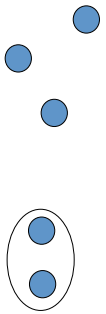

## Hierarchical clustering produces dramatic displays for expression profiles



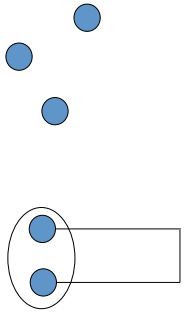

## Hierarchical clustering algorithm at work



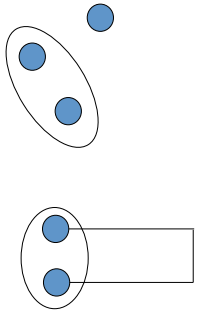
## Hierarchical clustering algorithm at work



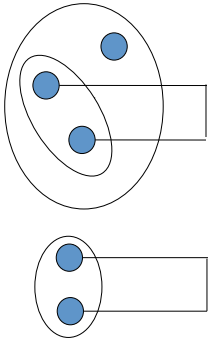
Hierarchical clustering algorithm at work



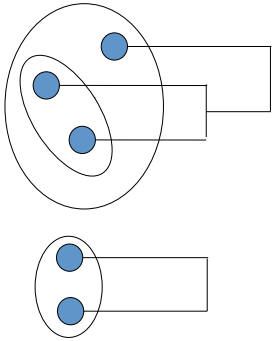
Hierarchical clustering algorithm at work



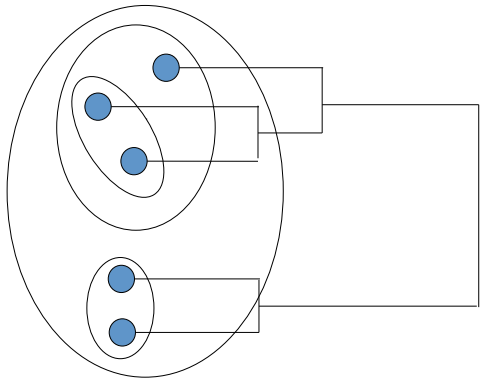
Hierarchical clustering algorithm at work



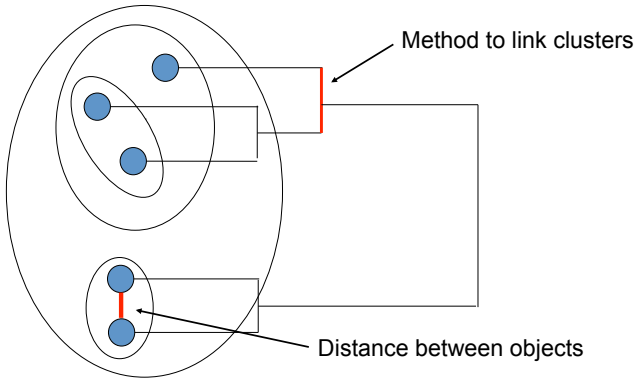
Hierarchical clustering algorithm at work



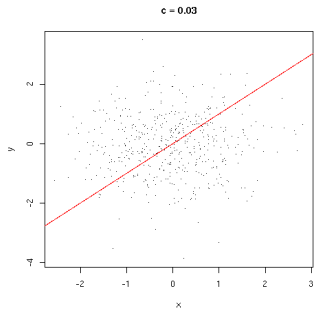
## Hierarchical clustering algorithm at work



## Hierarchical clustering comes in many (distance and linkage) flavors

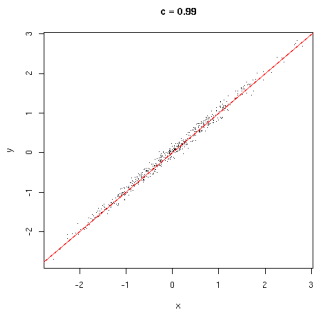



## Correlation is a popular distance for clustering microarray data



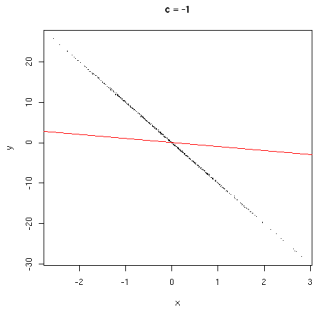

Unrelated profiles
have null a correlation

## Correlation is a popular distance for clustering microarray data



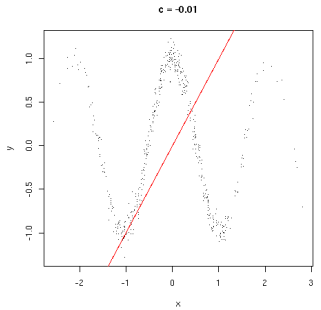

Similar profiles
have a unit correlation

## Correlation is a popular distance for clustering microarray data



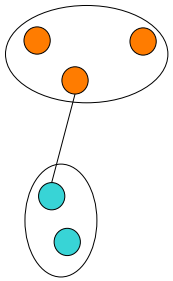

Correlation is scale invariant

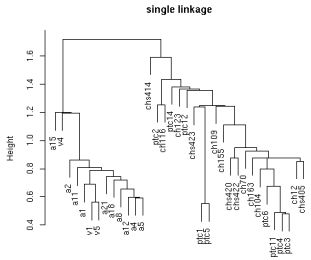
## Correlation is a popular distance for clustering microarray data



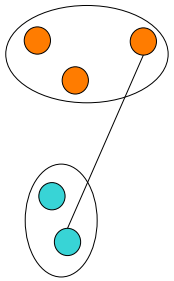

Correlation cannot detect nonlinear relations

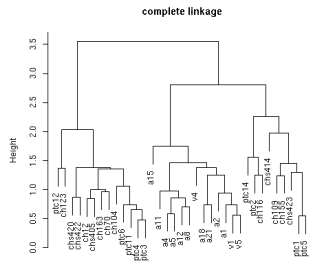
## Single linkage associates clusters with smallest minimal distance



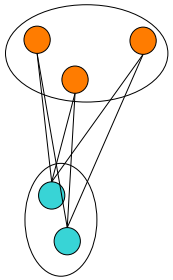


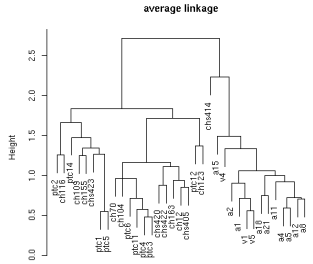

## Complete linkage associates clusters with smallest maximal distance



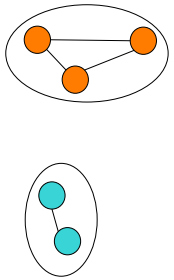


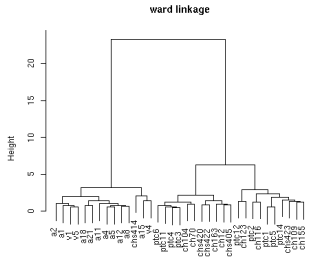
## Average linkage associates clusters with smallest average distance






## Ward linkage associates clusters as to minimizes within cluster variance





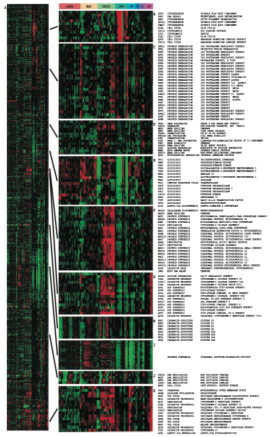

## Hierarchical clustering reveals structures in expression profiles

*Proc. Natl. Acad. Sci. USA*
Vol. 95, pp. 14863–14868, December 1998
Genetics

**Cluster analysis and display of genome-wide expression patterns**

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

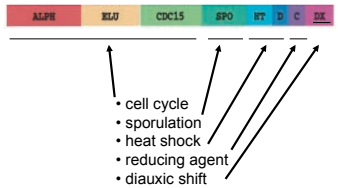
27



From Eisen et al., PNAS 1998

Yeast expression profiles were measured under various conditions

Two-ways hierarchical clustering was then applied to the expression matrix




28

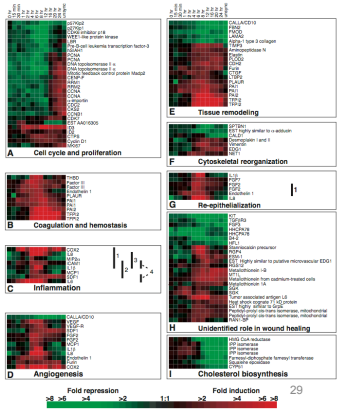
## Hierarchical clustering reveals the breadth of universal phenotypes

**The Transcriptional Program in the Response of Human Fibroblasts to Serum**

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

Iyer et al., Science 1999

- human fibroblasts were exposed to serum
- their expression profiles revealed a response akin to wound healing
- proliferaton is only part of this complex response




29

## Hierarchical clustering reveals molecular cancer subtypes

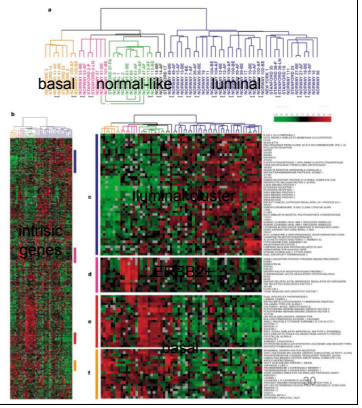
**letters to nature**

**Molecular portraits of human breast tumours**

Charles M. Perou*†‖, Therese Sørlie*‖, Michael B. Eisen*, Matt van de Rijn§, Stefanie S. Jeffrey*, Christian A. Rees*, Jonathan R. Pollack¶, Douglas T. Ross*, Hilde Johnsen‖, Lars A. Akslen#, Øystein Fluge**, Alexander Pergamenschikov*, Cheryl Williams*, Shirley X. Zhu‖, Per E. Lønning††, Anne-Lise Børresen-Dale‖, Patrick O. Brown*†† & David Botstein*

Perou et al., Nature 2000

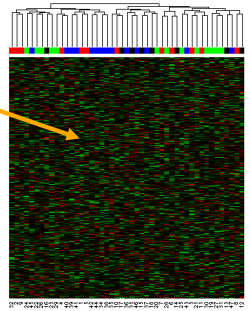
- clustering from 'intrisic', i.e. most variable, genes
- suggests that ER$^-$ tumors are *not* one but two diseases: basal-like and ERBB2$^+$




30

## Clustering produces clusters even when *no* clusters are there !!!

Random data



## Limits of hierarchical clustering

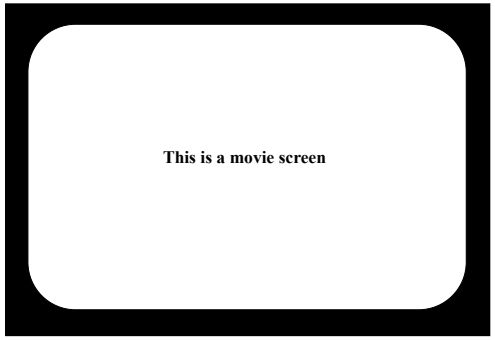
- ✓ It imposes a discrete structure on continuous data.
- ✓ Data do not always have a hierarchical structure.
- ✓ Classes may not be mutually exclusive
- ✓ These is no standard quality measure of how relevant a clustering is.
- ✓ It is unstable: adding one sample may incur dramatic changes in clustering pattern.

## Misuses of hierarchical clustering

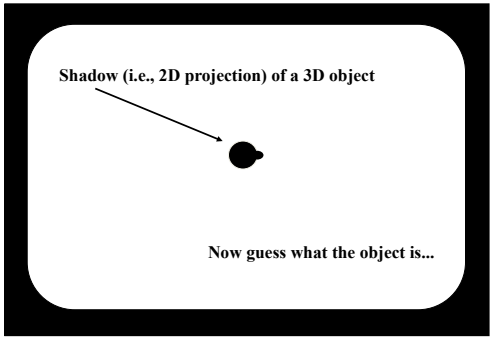
- ✓ It often used to (supposedly) assess signatures of classes which are known *a priori*. Supervised classification and cross-validation should be used instead.
- ✓ Clustering sometime relies on a small subset of predetermined class-separating genes. Such genes may be found by over-fitting any high dimension data set.

## Principal components analysis (PCA)

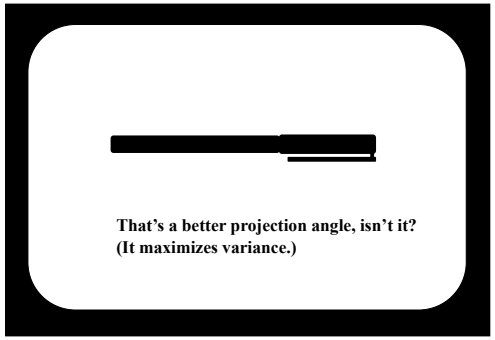
## Some projections of the data are more informative than others



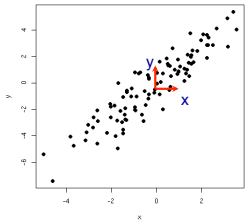

## Some projections of the data are more informative than others



Some projections of the data are more informative than others

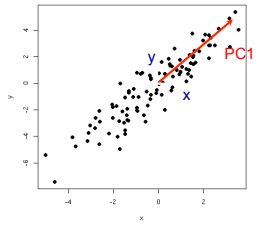



## Principal components analysis (PCA) computes the projection of the data that explains the best its variance




38

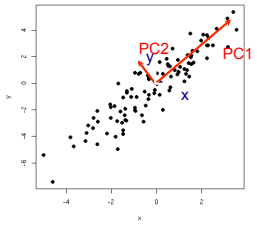
## Principal components analysis (PCA) computes the projection of the data that explains the best its variance




- PCA iteratively finds basis vectors (i.e. components) that maximise data variance

39

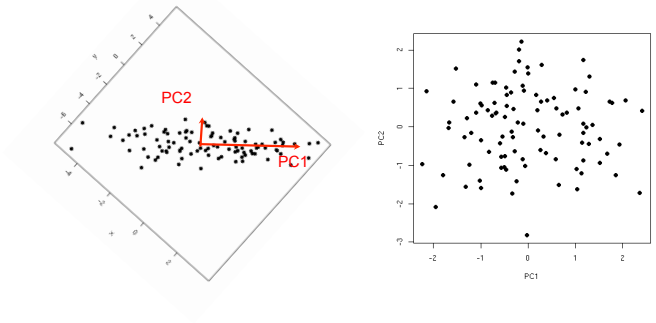
## Principal components analysis (PCA) computes the projection of the data that explains the best its variance




- PCA iteratively finds basis vectors (i.e. components) that maximise data variance
- Components are chosen orthogonal

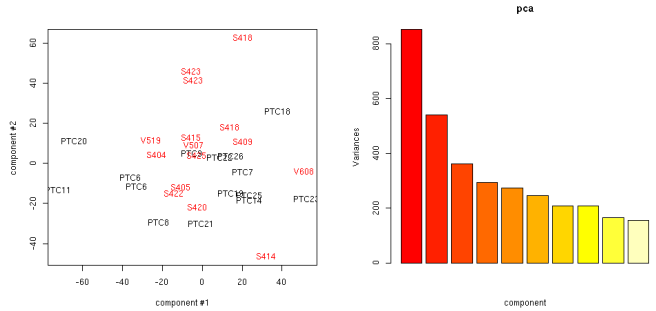
40

## Principal components analysis (PCA) computes the projection of the data that explains the best its variance

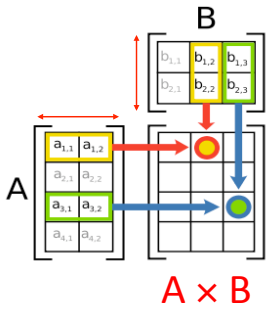



41

## Principal components analysis (PCA) computes the projection of the data that explains the best its variance




## Matrix product may represent any linear transformation




$(A \times B)_{1,2} = a_{1,1} \times b_{1,2} + a_{1,2} \times b_{2,2}$

$(A \times B)_{3,3} = a_{3,1} \times b_{1,3} + a_{3,2} \times b_{2,3}$

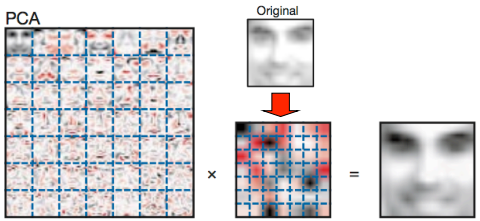
here the rank, $k$=2

A linear transformation is a combination of rotation, translation and stretching of individual base vectors

43

## Example of PCA application: compression of a faces database

Orthogonal 'face' basis encoding recovery




from Lee & Seung, Nature 1999

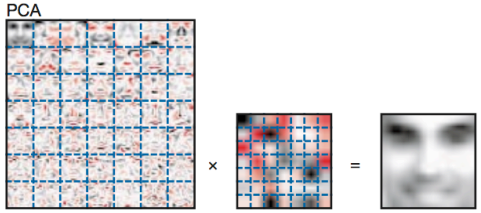
44

## Example of PCA application: compression of a faces database

transmit 49 'face' vectors

transmit 2,500 encoding vectors

PCA



from Lee & Seung, Nature 1999

45

## Example of PCA application: compression of a faces database

- Size of image database = 19x19x2500
- Size of PCA-compressed database = size of base vectors + size of faces encodings = 19x19x49 + 49x2500
- Compression ratio about 6.4

46

## What works for faces works for expression matrices

$A = W \times H$

faces database = face basis × face encodings

expression matrix = sample basis × sample encodings

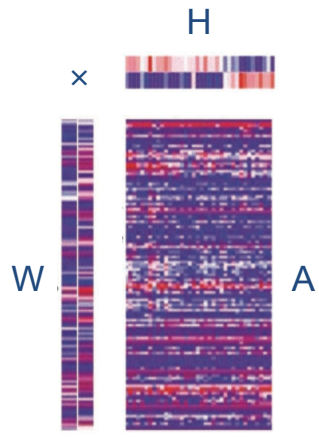
47

## $A \approx W \times H$

A: aproximate expression matrix

W: sample basis

H: sample encoding in term of k=2 'super genes'




48

## How variable is mRNA expression among ethnic groups?

Let's have a look at the data from this paper:

Common genetic variants account for differences in gene expression among ethnic groups

Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]

Nature Genetics, 2007

49

## Why is the question important?

- It may help in the design of clinical trails and health policies by identifying distinct disease susceptibility and/or drug response profiles in specific populations
- It is underlying the origin and diversity of humans
- It is politically loaded by an *unthinkable* history of racism and ethnic/racial violence across the world, some of it very recent and close to home

50

## The populations investigated by Spielman et al.

166 individuals were profiled

- 60 Caucasian Europeans
- 41 Japanese
- 41 Chinese living in Beijin, China

HapMap individuals

- 24 Chinese living in Los Angeles, USA. These are controls for the effect of environment and life-style

51

## The data of Spielman et al.

- mRNA was extracted from Epstein-Barr virus-transformed lymphoblastoid cell lines
- Expression was measured with Affymetrix HG Focus arrays covering ~8,500 genes
- They are publicly available from www.ncbi.nlm.nih,gov/GEO, accession number GSE5859

52

## Follow up: pervasive interethnic difference or batch effect?

A few months later Akey et al. (Nature Genet. 2007) revisited Cheung's study:

- "To explore these issues in more detail, we downloaded the raw CEL files from Gene Expression Omnibus (GSE5859) and extracted from the header line the date on which the file was created."
- "Interestingly, the arrays used to measure expression for the CEU individuals were primarily processed from 2003 to 2004, whereas the arrays used to measure expression for the ASN individuals were all processed in 2005–2006."

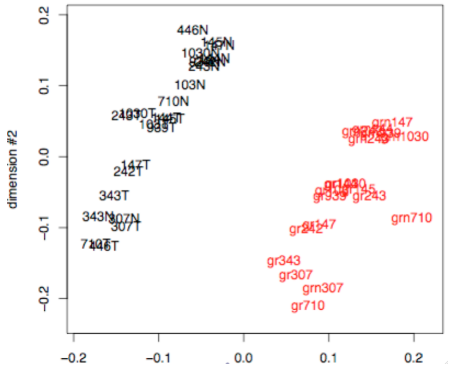
53

## Follow up: pervasive interethnic difference or batch effect?

- 94% the genes associated with batch
- No differential expression associated with ethnicity after removal of batch effects
- Batch is a *confounder*: one cannot conclude whether ethnicity or batch explains Cheung's observations
- Further comparison of dates within the CEU samples shows that at least 79% of the genes are still associated with batch

54

## Batch effects are often the *main* determinant of gene expression




55

## Batch effects are often the *main* determinant of gene expression

- Randomize all experimental steps
- Beware of cross-study data merging
- Removing batch effects as the potential to remove key biological signal in the data or introduce more spurious signal
- […Still, Storey et al. (American J. Hum. Genet, 2007) report that 17% of the genes were differentially expressed between a group of African and Europeans]

56

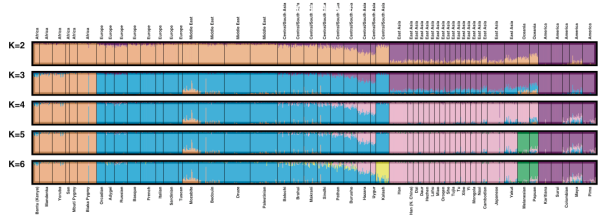
# Probing the geographic heterogeinty of human populations

377 microsatellites genotyped in 1056 individuals from 52 populations

k-mean clustering was applied to the data

between-population accounts for 3-5% of total variance

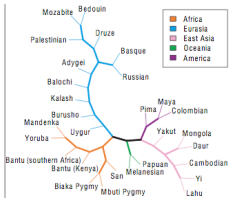
3-5% is enough to accurately classify individuals according to their origin



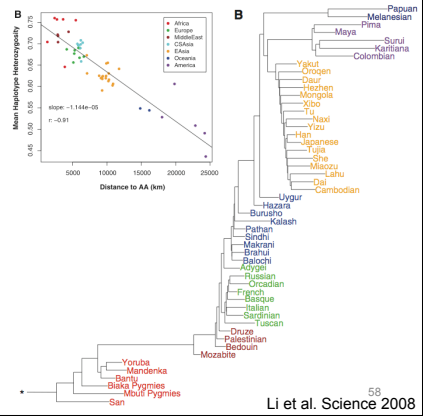

Rosenberg et al. Science 2002



# Reconstructing with SNP arrays the history of human populations




Jakobsson et al., Nature 2008






Li et al. Science 2008

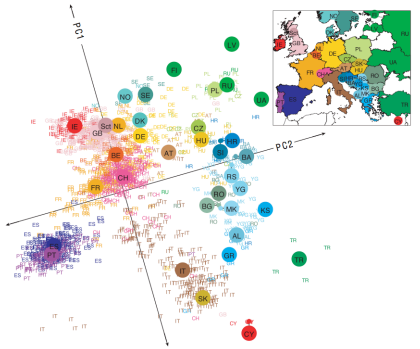
# Genes mirror geography within Europe

Novembre et al. Nature, 2008

- ~3200 Eurpeans were genotyped with 500K Affymetrix SNP arrays
- Individual with non European ancestry or with discordant grand-parental origins were removed
- Data were analysed with PCA



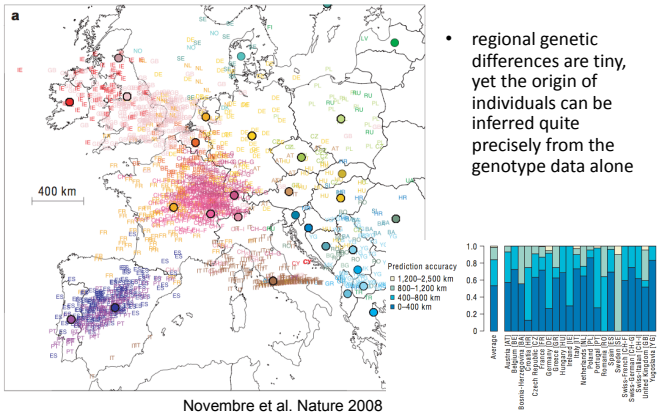
# Genes mirror geography within Europe




- PC1 is aligned along a North-South axis
- PC2 is aligned along a East-West axis
- PC1 and PC2 account for 0.5% of total variance
- Thus, 99.5% of the variance is *not* related to geography

Novembre et al. Nature 2008

## Genes mirror geography within Europe




- regional genetic differences are tiny, yet the origin of individuals can be inferred quite precisely from the genotype data alone

Novembre et al. Nature 2008

61

## Discrimination always starts with a definition of the victims…

nature
www.nature.com/nature Vol 461 | Issue no. 7265 | 8 October 2009

**Genetics without borders**

A UK government scheme to establish nationality through DNA testing is scientifically flawed, ethically dubious and potentially damaging to science.

So it was with understandable incredulity that researchers received a plan by the UK Border Agency to use genetics to determine nationality — specifically, the origin of asylum-seekers claiming to be from war-torn Somalia. The agency's pilot programme, which began last month, aims to determine whether some 100 individuals really are Somali nationals by checking them for the individual DNA variants known as single nucleotide polymorphisms (SNPs) in mitochondrial DNA, on the Y chromosome and elsewhere in the genome. The scheme will also use isotopic ratios of elements found in hair and fingernails — which can vary depending on a person's diet or environment — to try to establish where the migrants previously lived.

62

## Strengths and limits of principal components analysis

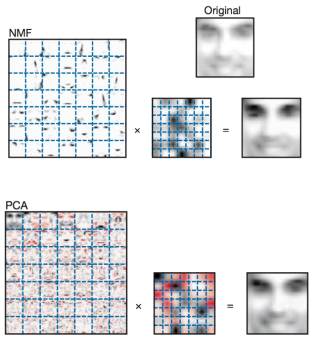
- ✓ The percentage of variance explained by the first components is a good quality measure.
- ✓ It produces continuous representations.
- ✓ Conceptual simplicity, and strong mathematical theory.

But:

- ✓ Components do not always have simple biological interpretations.
- ✓ No choice of distance measure.

## Nonnegative matrix factorization (NMF)

## NMF represents the data in term of nonnegative components




- The nonnegativity constrainst yields components representing *parts of objects* (here nose, eye, mouth...), which are interpretable
- NMF components are *not* orthogonal

from Lee & Seung, Nature 1999

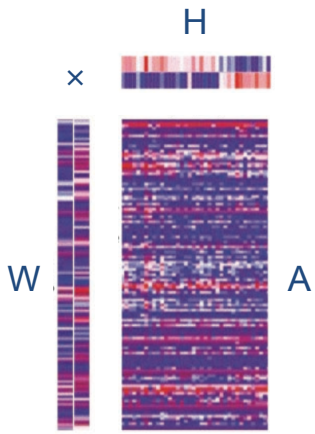


## $A \approx W \times H$

A: aproximate expression matrix
H: metagenes
W: $w_{ij}$ is the coefficient of gene i in metagene j

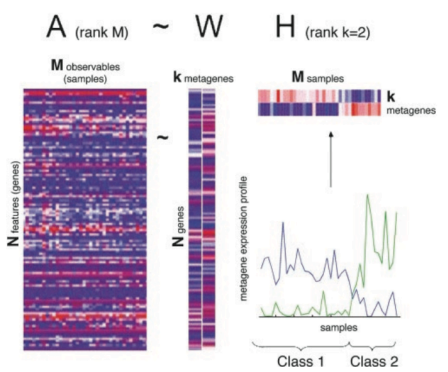
$w_{ij}>0$, $h_{ij}>0$

metagenes are *not* othogonal a priori






## Metagenes split the samples in clusters, but do not force a hierarchy




k is an adjustable parameter

from Brunet et al., PNAS 2004



## NMF is a stochastic algorithm

- NMF runs from different initial conditions may converges to different local mimina
- If they are k clearly defined clusters, one expect most rank k NMF runs to reveal metagenes mirroring these k clusters

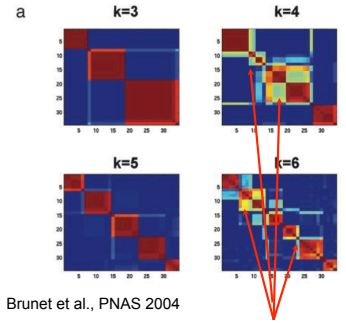
**Idea:** use the consistency of NMF results across sets of random initial condition may be used as a measure of clustering quality

## The consensus matrix represents the consistency of sample assignment to clusters

- For each run compute a *connectivity matrix* MxM matrix, C, with $c_{ij}$=1 if samples i and j are in the same cluster, $c_{ij}$=0 otherwise
- The average of connectivity matrices over many runs, |C| is called the *consensus matrix*
- One can reorder the rows and columns of |C| with average linkage hierarchical clustering

69

## The consensus matrix represents the consistency of sample assignment to clusters



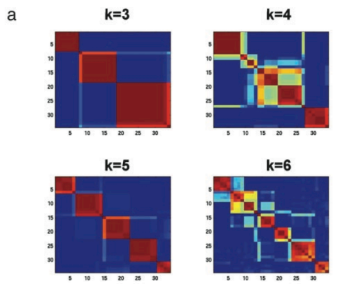

Brunet et al., PNAS 2004

- 25 classic and 9 desmoplastic medulloblastoma mRNA profiles (~6,000 genes)
- Consensus of 50 NMF runs are shown for k in {3,… 6}
- k=2 also yields clear-cut classes (not shown)

unstable assignments

70

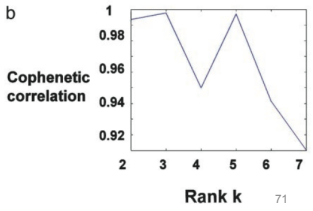
## The cophenetic correlation summarizes the dispersion of the consensus matrix




Brunet et al., PNAS 2004

$$c = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{[\sum_{i<j}(x(i,j)-x)^2][\sum_{i<j}(t(i,j)-t)^2]}}.$$

x(i, j) = distance between i and j
t(i, j) = distance between i and j in consensus matrix dendrogram
x is the average of x(i, j), t of t(i, j)




71

## Other extentions of PCA have been proposed

- NMF uses nonnegative matrix
- ICA uses independent components
- kernel PCA maps data in higher dimension first (using SVM's "kernel trick")

72