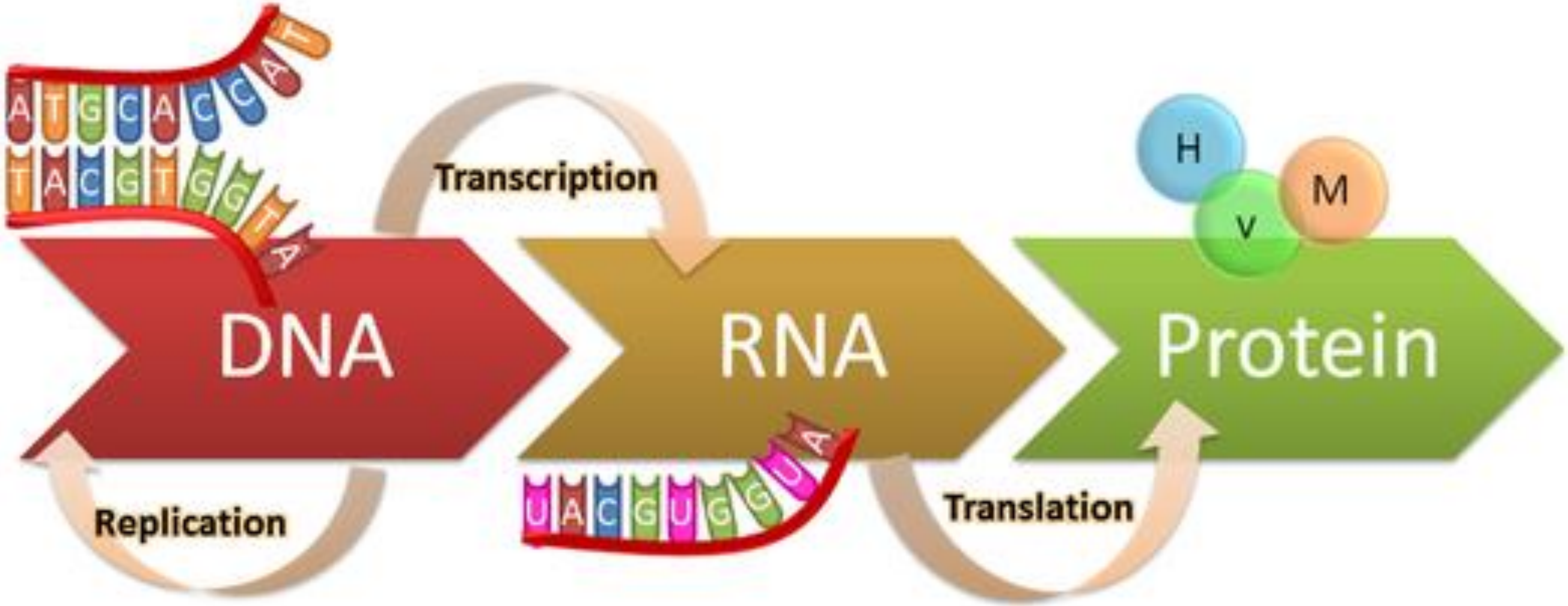
Matlab Homework 7: Microarray Gene Expression

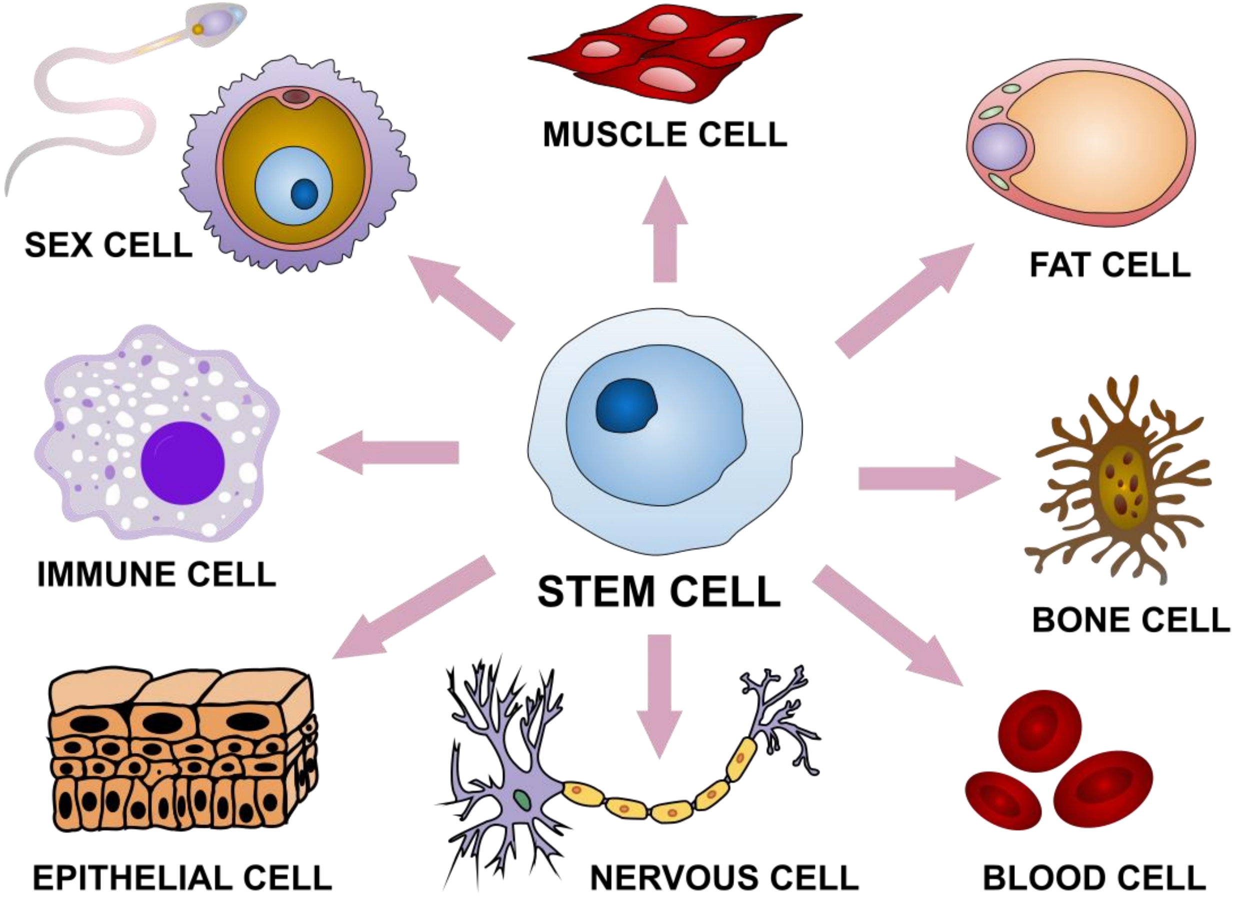


- **In genetics, gene expression (genotype) gives rise to the phenotype, i.e. observable trait.**
- **Mistakes in gene expression result in diseases in human**

https://genius.com/Biology-genius-the-central-dogma-annotated

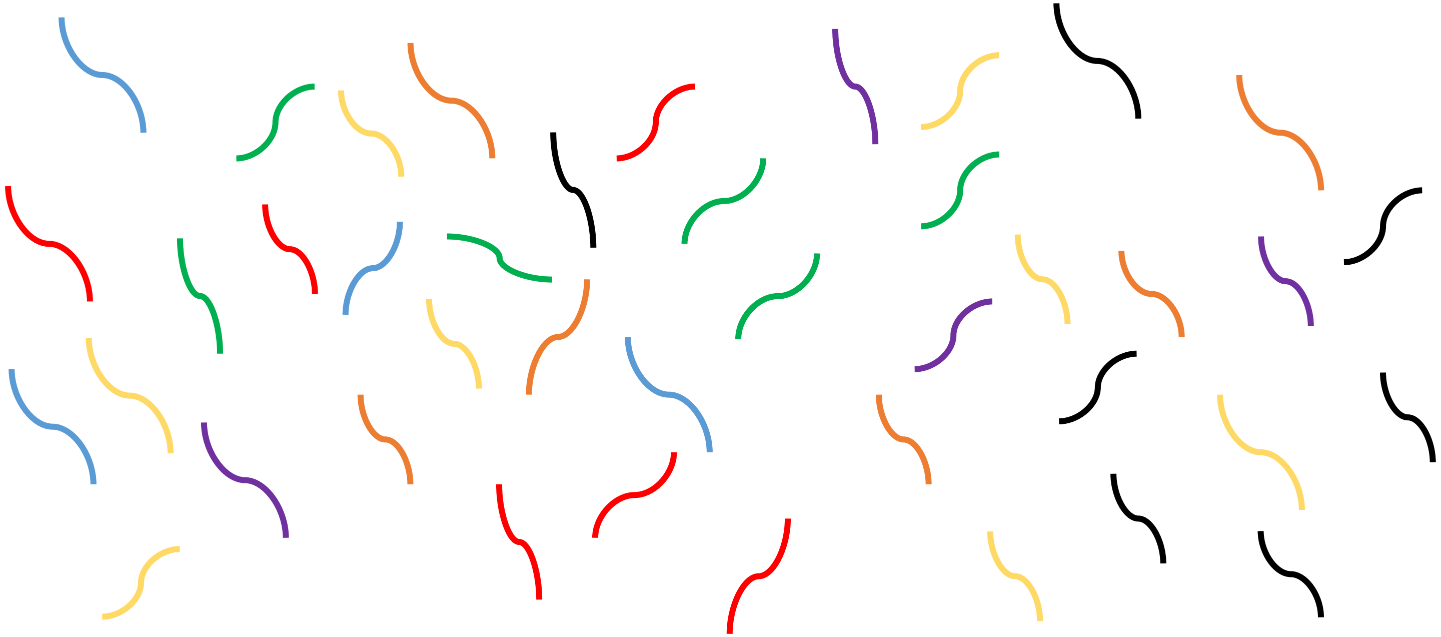
Differential gene expression (DGE)

- **Spatio-temporal gene expression of 20,000 genes**
- **Which genes are differentially expressed in the two different tissue types**
- **Which genes are differentially expressed in normal tissue and disease tissue**

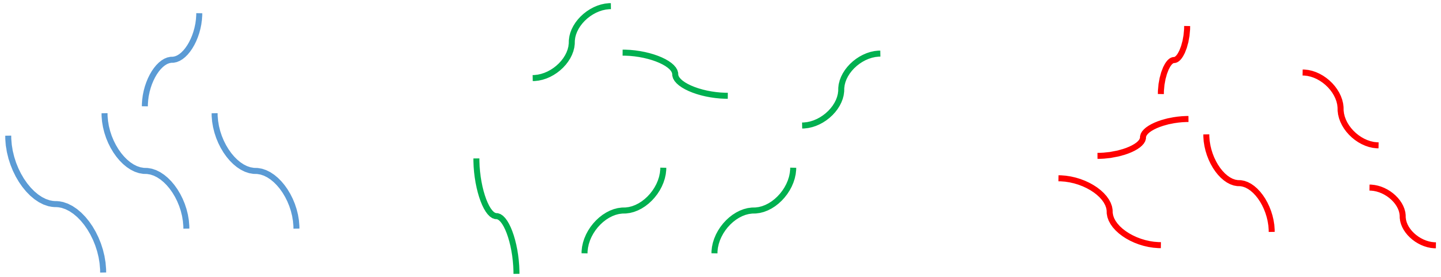


Quantify the gene expression

Mixed mRNA in 20,000 genes



Sort them and Measure the individual gene expression



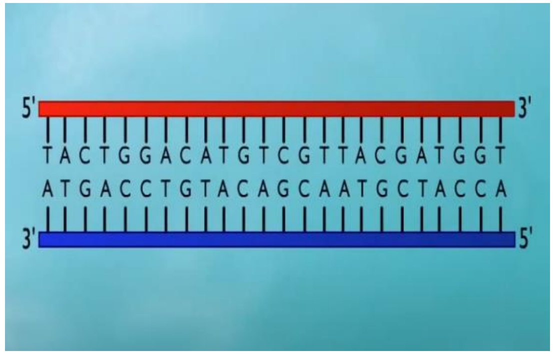
Ways to quantify the gene expression

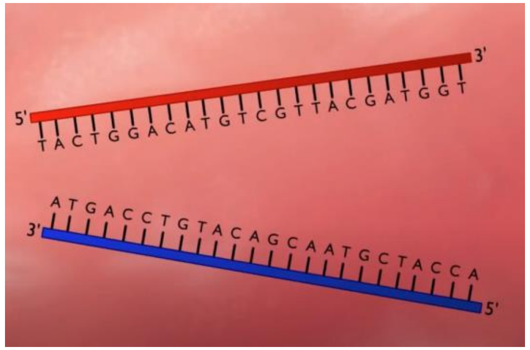
- Northern/Western blotting
- Fluorescent in situ hybridization (FISH)
- Reverse transcription polymerase-chain-reaction (RT-PCR)
- Serial Analysis of Gene Expression (SAGE)
- DNA microarray (high throughput)
- RNA Sequencing (high throughput)

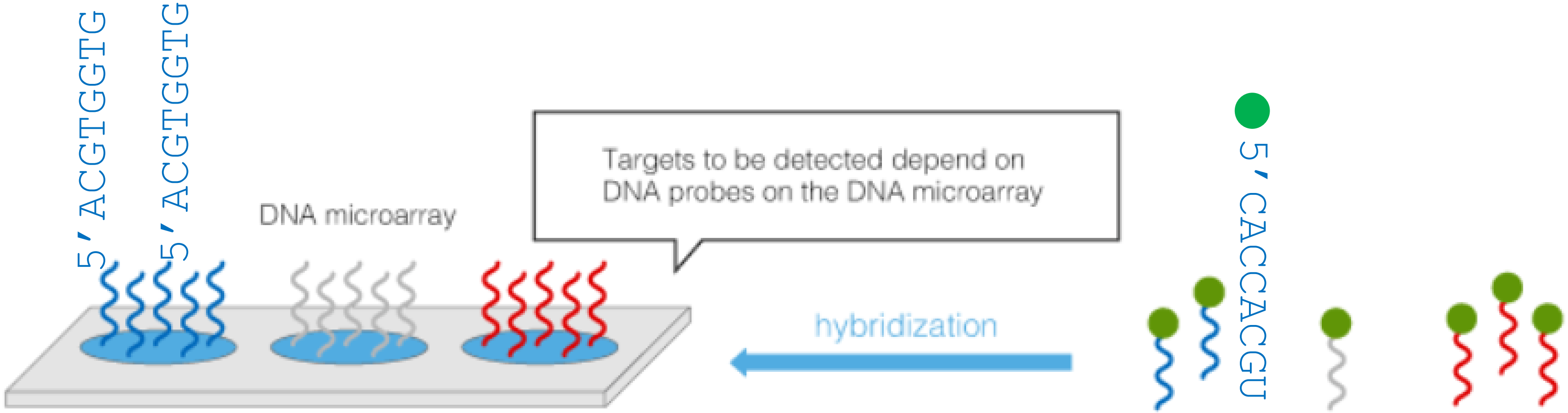
DNA microarray (high throughput) is based on the principle of DNA/RNA hybridization

Double-stranded DNA

Positive/forward strand
Reverse/complementary strand







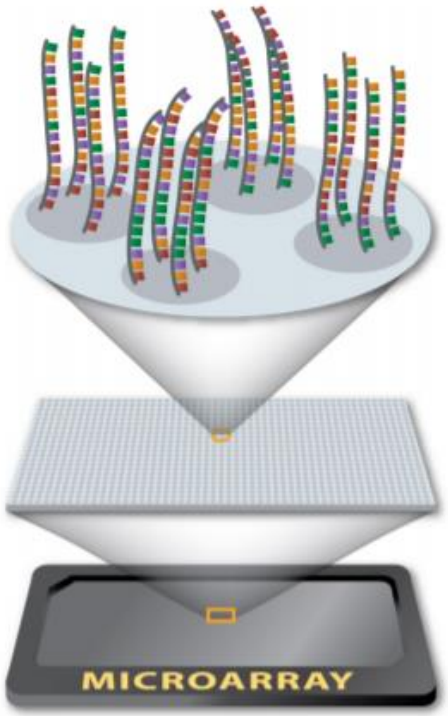
Targets to be detected depend on DNA probes on the DNA microarray

Gene1 Gene2 Gene3

Attach green fluorophores to the mRNA molecules

gene expression: which gene is expressed and how much (fluorescence intensity)

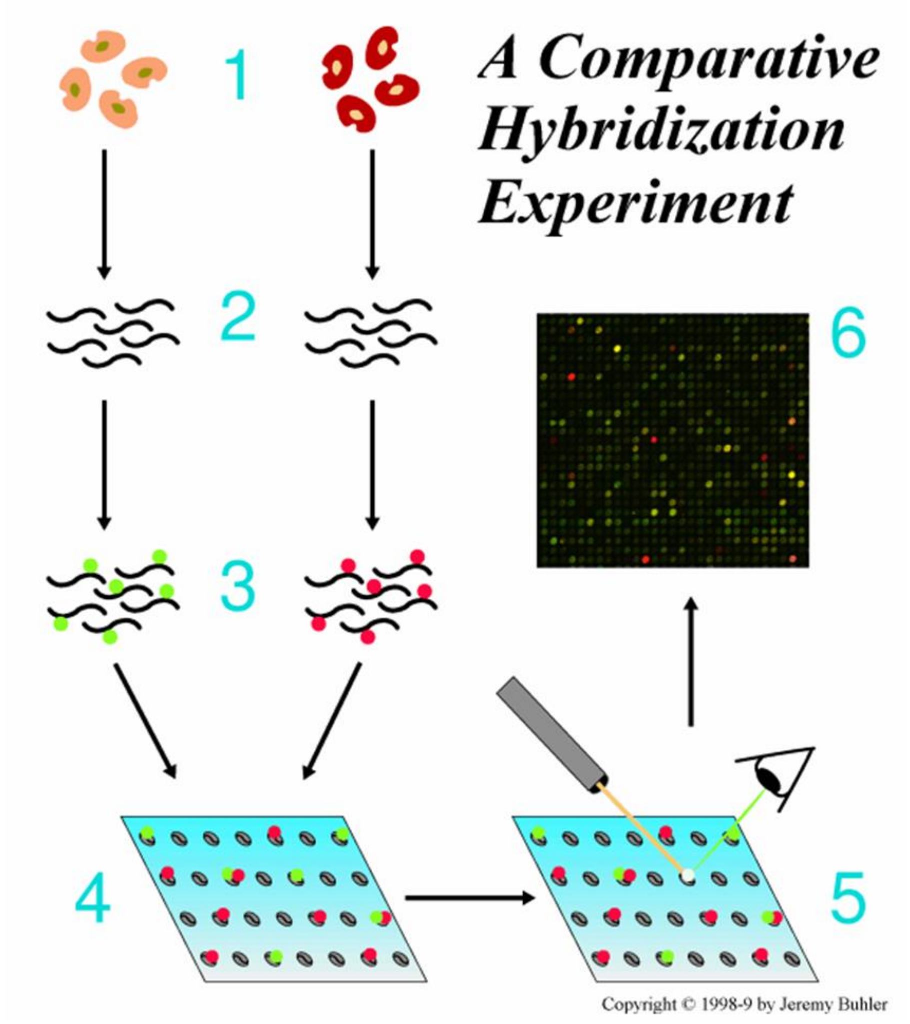
DNA microarray Fabrication



http://learn.genetics.utah.edu

A typical comparative microarray experiment

Normal tissue Cancer tissue



Copyright © 1998-9 by Jeremy Buhler

Extract RNA

Convert RNA to cDNA
and label them with green/red
color dye dUPT-cy3

Mix and hybridize on the array

Scan and image

Analyze the data

http://www.bio.davidson.edu/Biology/Courses/genomics/chip/chip.html

Gene expression data analysis

- Data normalization
- Clustering and Classification
- Databases and software

Gene expression data example

Data on *m* genes for *n* samples

ID_REF	**GSM136326**	**GSM136327**	**GSM136328**	**GSM136329**	**GSM136330**	**GSM136331**
1007_s_at	10.4502763	9.3995422	9.42479936	9.472922422	9.27878032	9.434427931
1053_at	5.71946574	4.84929333	4.73208086	4.728854347	5.32639216	5.230320408
117_at	5.93866366	6.08327317	6.44797814	6.17694869	6.54458475	6.07779478
121_at	8.0230524	7.8946588	8.34498775	8.163203547	8.23375629	7.595105829
1255_g_at	3.95480312	3.96324647	3.96410203	4.087835849	3.99889298	3.839704814
1294_at	7.9090045	8.36397325	8.27191179	8.358196216	7.69999823	8.274305066
1316_at	6.50101269	7.06478465	6.84193046	7.16941856	6.47412502	6.182111503
1320_at	4.46782927	4.44699592	4.55541274	4.660870385	4.74765838	4.484930349
1405_i_at	6.98704598	6.91061747	6.70850464	7.777802579	7.74606814	7.576616437
1431_at	3.78442846	3.84921957	3.86910251	4.06116236	3.90178173	3.669844515
1438_at	6.33719475	6.35968637	6.26322624	6.571037625	6.41512718	5.992213616
1487_at	7.6496267	7.16653011	7.25328005	7.228355117	7.14035537	7.626439717

Gene expression data analysis-analyze each gene

ID_REF	GSM136326	GSM136327	GSM136328	GSM136329	GSM136330	GSM136331
1007_s_at	10.4502763	9.3995422	9.42479936	9.472922422	9.27878032	9.434427931
1053_at	5.71946574	4.84929333	4.73208086	4.728854347	5.32639216	5.230320408

Group the patients into difference groups and see if there is a difference between them for each gene

Group 1

ID_REF	GSM136326	GSM136327
1007_s_at	10.4502763	9.3995422
1053_at	5.71946574	4.84929333

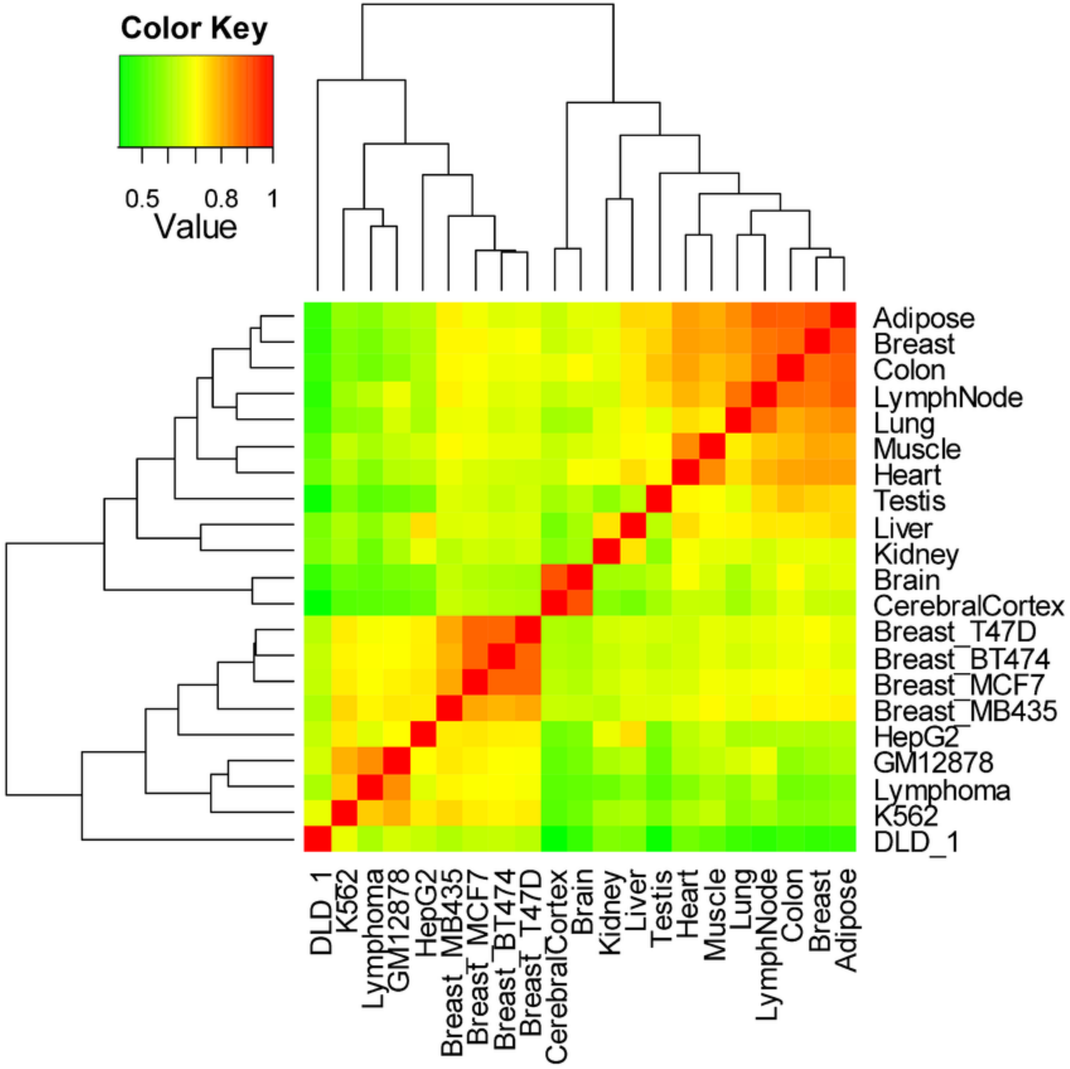
Group 2

ID_REF	GSM136328	GSM136329	GSM136330	GSM136331
1007_s_at	9.42479936	9.472922422	9.27878032	9.434427931
1053_at	4.73208086	4.728854347	5.32639216	5.230320408

Perform ttest for 1007_s_at p=0.0001
Perform ttest for 1007_s_at p=0.0002

Find the fold change.
Average gene expression of 1007_s_at (group 1)/ Average geneexpression of 1007_s_at (group 2)

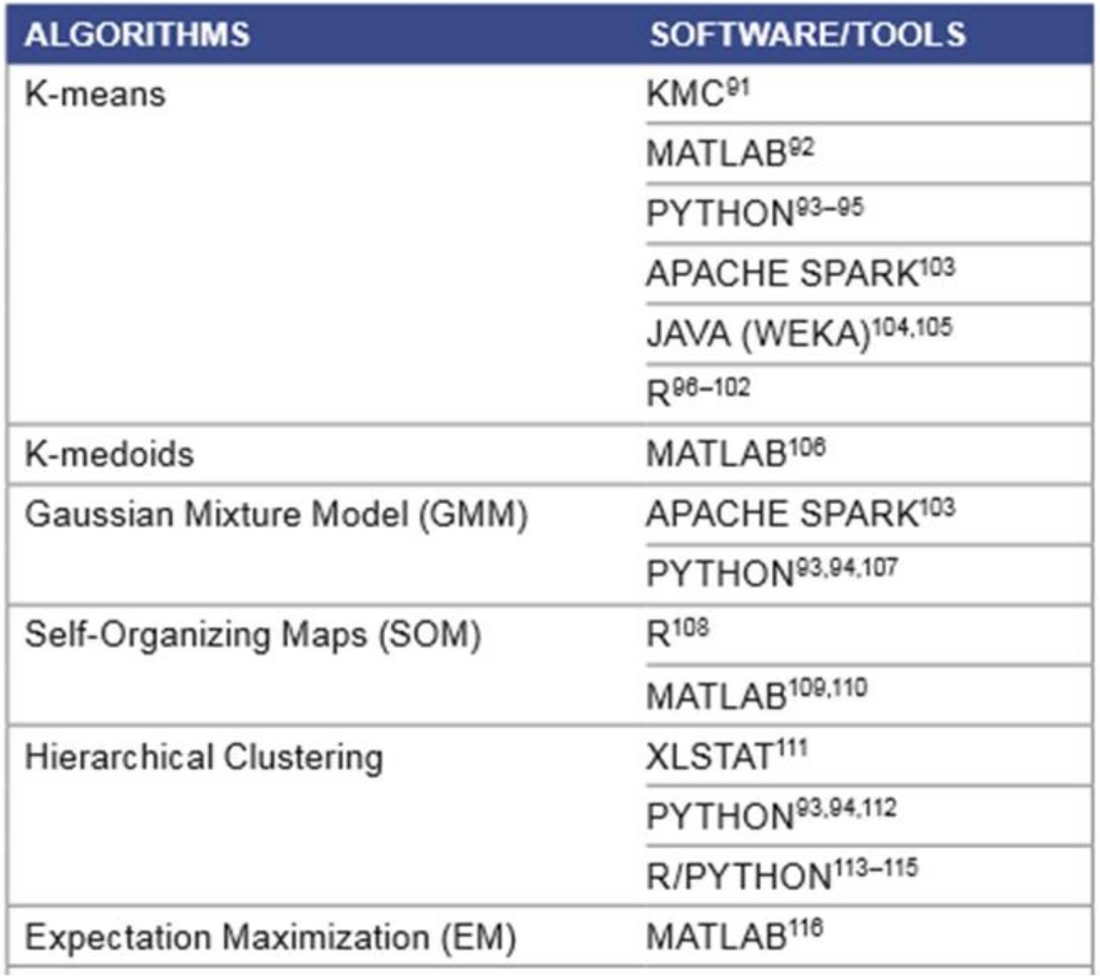
Gene Clustering



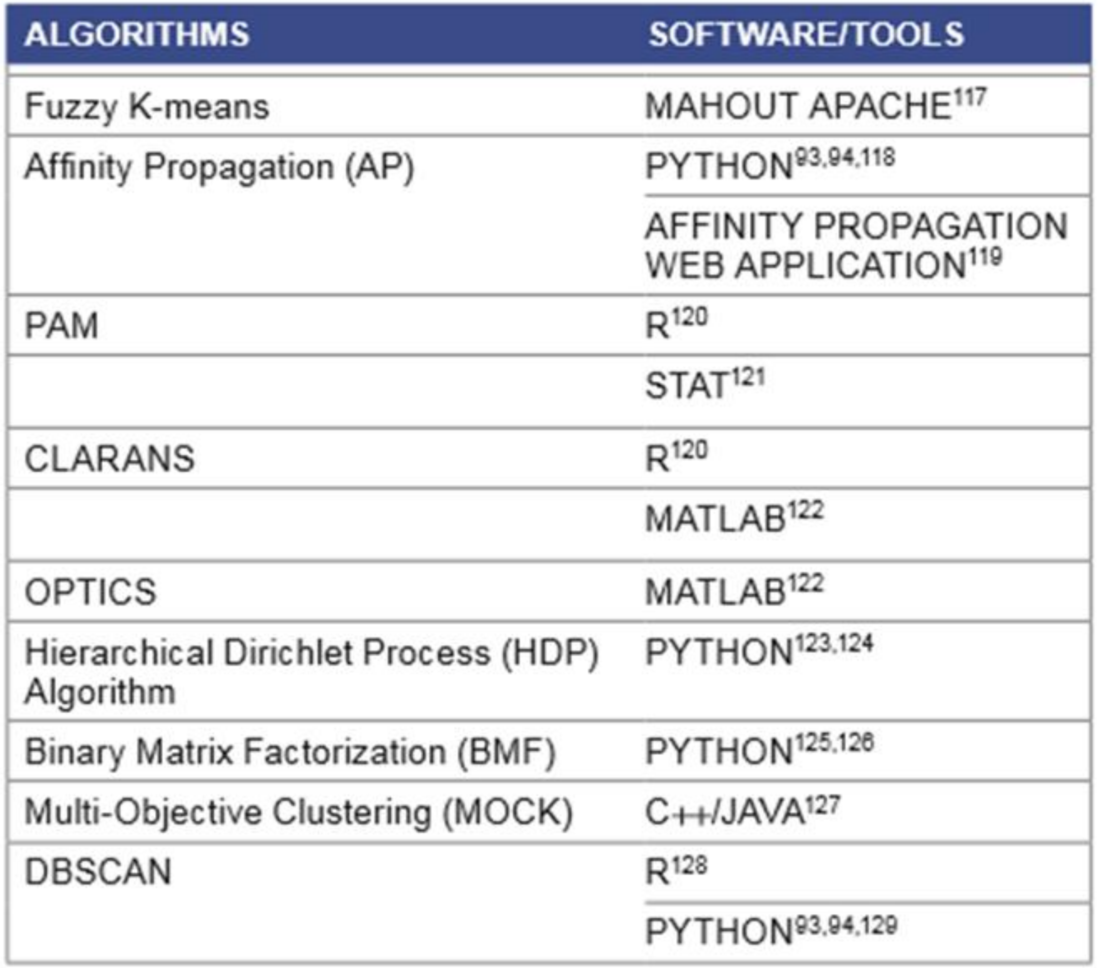
https://www.researchgate.net/figure/Gene-expression-profiles-hierarchical-cluster-analysis-based-on-Spearman_fig22_235402788

Databases and software

Deposit data in the public database(http://www.ncbi.nlm.nih.gov/geo/)



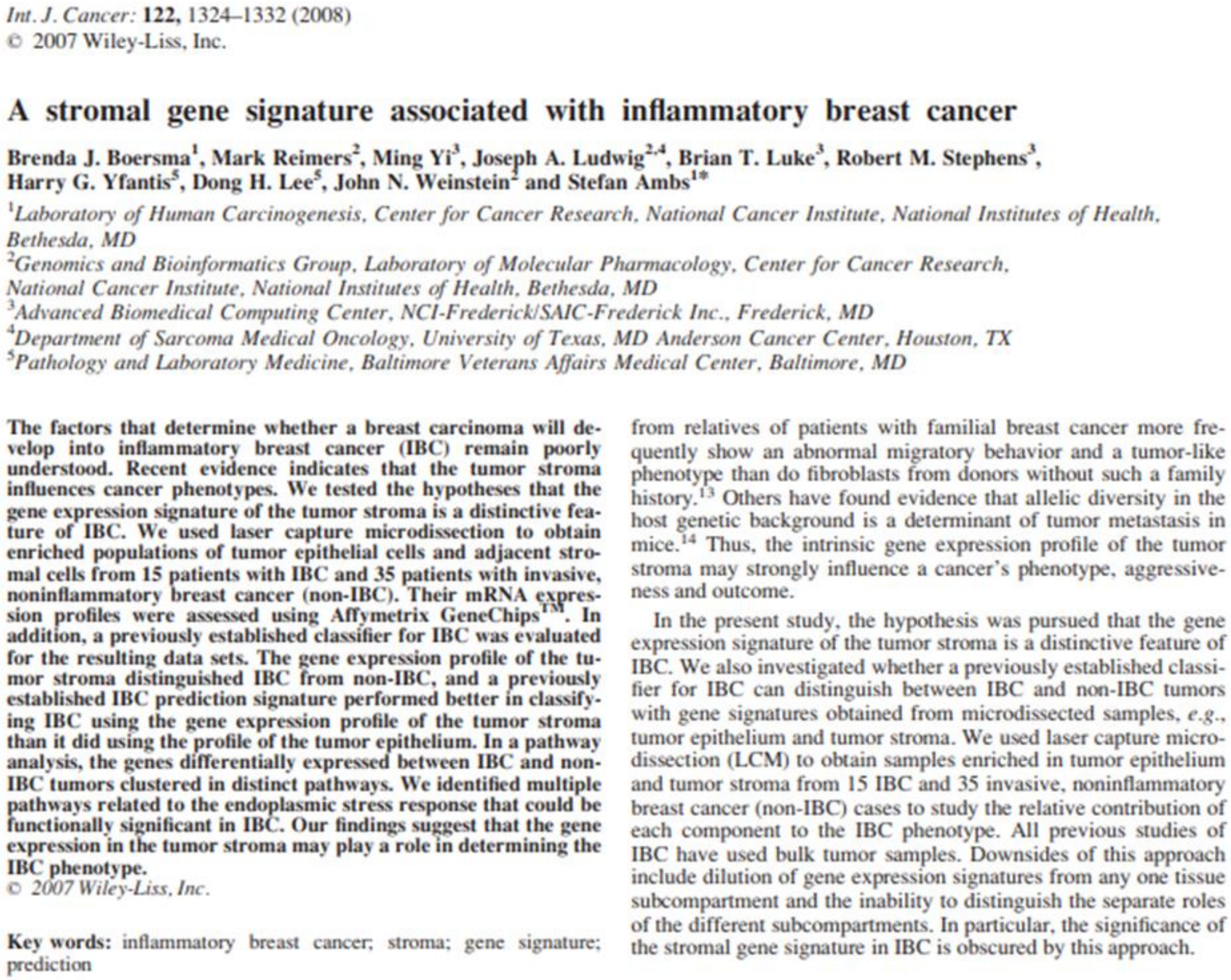
ALGORITHMS	SOFTWARE/TOOLS
K-means	KMC[91]
	MATLAB[92]
	PYTHON[93–95]
	APACHE SPARK[103]
	JAVA (WEKA)[104,105]
	R[96–102]
K-medoids	MATLAB[106]
Gaussian Mixture Model (GMM)	APACHE SPARK[103]
	PYTHON[93,94,107]
Self-Organizing Maps (SOM)	R[108]
	MATLAB[109,110]
Hierarchical Clustering	XLSTAT[111]
	PYTHON[93,94,112]
	R/PYTHON[113–115]
Expectation Maximization (EM)	MATLAB[116]



ALGORITHMS	SOFTWARE/TOOLS
Fuzzy K-means	MAHOUT APACHE[117]
Affinity Propagation (AP)	PYTHON[93,94,118]
	AFFINITY PROPAGATION WEB APPLICATION[119]
PAM	R[120]
	STAT[121]
CLARANS	R[120]
	MATLAB[122]
OPTICS	MATLAB[122]
Hierarchical Dirichlet Process (HDP) Algorithm	PYTHON[123,124]
Binary Matrix Factorization (BMF)	PYTHON[125,126]
Multi-Objective Clustering (MOCK)	C++/JAVA[127]
DBSCAN	R[128]
	PYTHON[93,94,129]

Oyelade, J. et al. (2016) Clustering algorithms: their application to gene expression data. Bioinformatics and Biology insights, 10, BBI. S38316.

IBC vs. non-IBC
Stromal vs. epithelial



Int. J. Cancer: **122,** 1324–1332 (2008)

A stromal gene signature associated with inflammatory breast cancer

Brenda J. Boersma[1], Mark Reimers[2], Ming Yi[3], Joseph A. Ludwig[2,4], Brian T. Luke[3], Robert M. Stephens[3], Harry G. Yfantis[5], Dong H. Lee[5], John N. Weinstein[2] and Stefan Ambs[1]*

[1]*Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD*

[2]*Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD*

[3]*Advanced Biomedical Computing Center, NCI-Frederick/SAIC-Frederick Inc., Frederick, MD*

[4]*Department of Sarcoma Medical Oncology, University of Texas, MD Anderson Cancer Center, Houston, TX*

[5]*Pathology and Laboratory Medicine, Baltimore Veterans Affairs Medical Center, Baltimore, MD*

The factors that determine whether a breast carcinoma will develop into inflammatory breast cancer (IBC) remain poorly understood. Recent evidence indicates that the tumor stroma influences cancer phenotypes. We tested the hypotheses that the gene expression signature of the tumor stroma is a distinctive feature of IBC. We used laser capture microdissection to obtain enriched populations of tumor epithelial cells and adjacent stromal cells from 15 patients with IBC and 35 patients with invasive, noninflammatory breast cancer (non-IBC). Their mRNA expression profiles were assessed using Affymetrix GeneChips™. In addition, a previously established classifier for IBC was evaluated for the resulting data sets. The gene expression profile of the tumor stroma distinguished IBC from non-IBC, and a previously established IBC prediction signature performed better in classifying IBC using the gene expression profile of the tumor stroma than it did using the profile of the tumor epithelium. In a pathway analysis, the genes differentially expressed between IBC and non-IBC tumors clustered in distinct pathways. We identified multiple pathways related to the endoplasmic stress response that could be functionally significant in IBC. Our findings suggest that the gene expression in the tumor stroma may play a role in determining the IBC phenotype.
© 2007 Wiley-Liss, Inc.

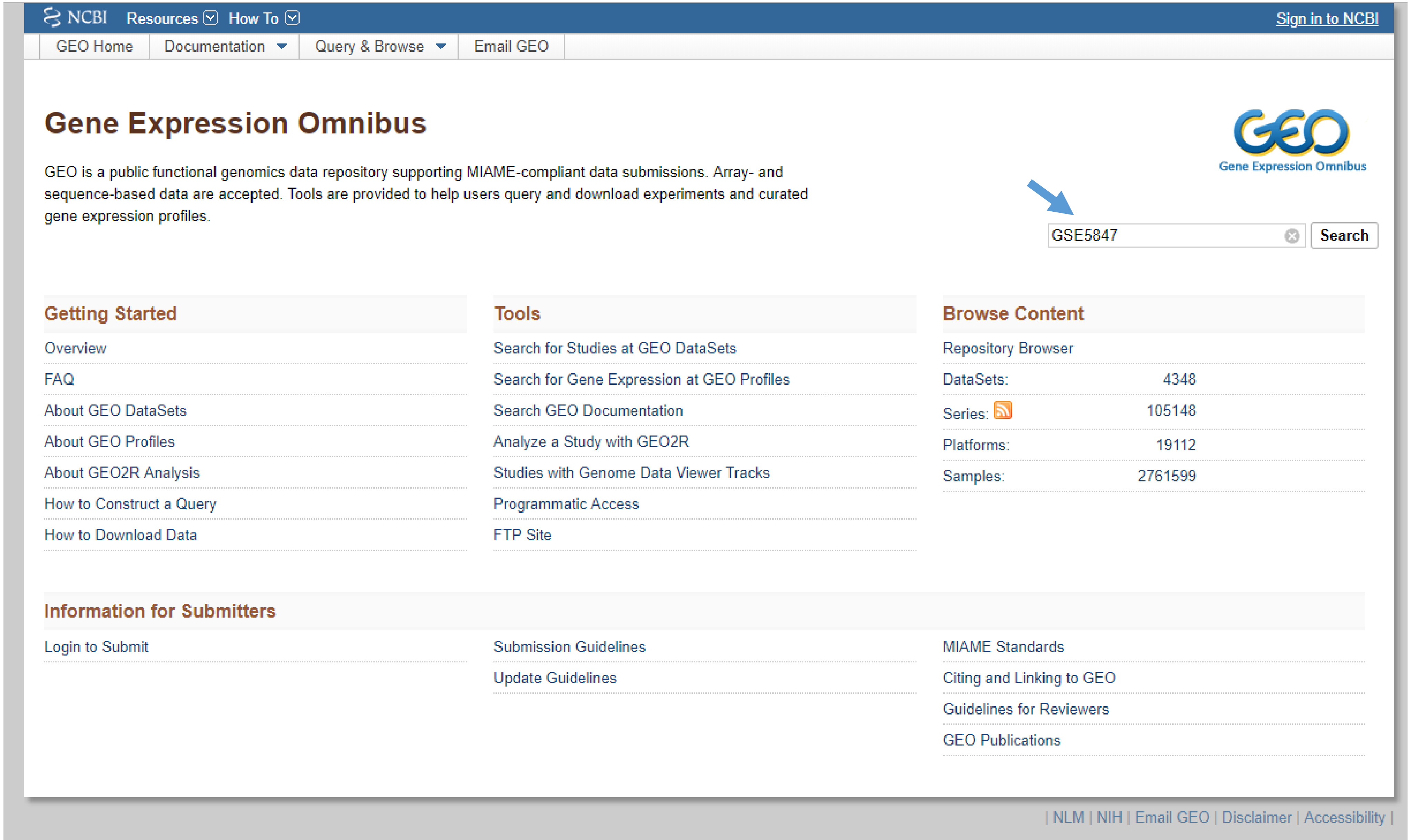
Key words: inflammatory breast cancer; stroma; gene signature; prediction

from relatives of patients with familial breast cancer more frequently show an abnormal migratory behavior and a tumor-like phenotype than do fibroblasts from donors without such a family history.[13] Others have found evidence that allelic diversity in the host genetic background is a determinant of tumor metastasis in mice.[14] Thus, the intrinsic gene expression profile of the tumor stroma may strongly influence a cancer's phenotype, aggressiveness and outcome.

In the present study, the hypothesis was pursued that the gene expression signature of the tumor stroma is a distinctive feature of IBC. We also investigated whether a previously established classifier for IBC can distinguish between IBC and non-IBC tumors with gene signatures obtained from microdissected samples, *e.g.*, tumor epithelium and tumor stroma. We used laser capture microdissection (LCM) to obtain samples enriched in tumor epithelium and tumor stroma from 15 IBC and 35 invasive, noninflammatory breast cancer (non-IBC) cases to study the relative contribution of each component to the IBC phenotype. All previous studies of IBC have used bulk tumor samples. Downsides of this approach include dilution of gene expression signatures from any one tissue subcompartment and the inability to distinguish the separate roles of the different subcompartments. In particular, the significance of the stromal gene signature in IBC is obscured by this approach.

Deposit data in the public database
(http://www.ncbi.nlm.nih.gov/geo/)

- **Data description:**(http://www.ncbi.nlm.nih.gov/geo/)



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

Repository Browser

DataSets:	4348
Series:	105148
Platforms:	19112
Samples:	2761599

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

| NLM | NIH | Email GEO | Disclaimer | Accessibility |

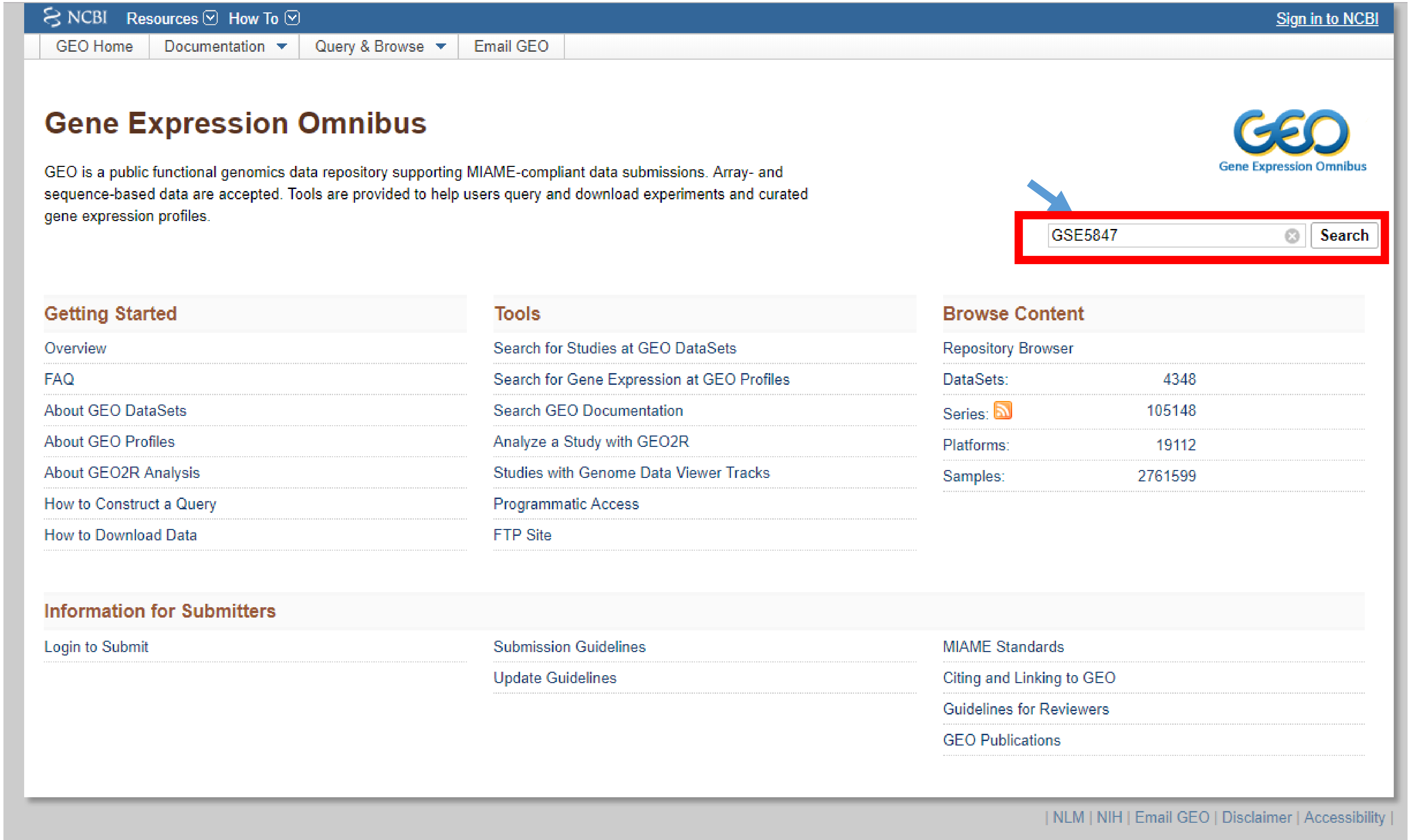
GEO data types

- GSM: An individual microarray sample (e.g., patient, or tissue).

- GSE: A Series, representing an experimental study. A GSE contains one or more GSM entries.
Probe-GSM-data(intensity ratio)

- GPL: Platform data, containing information about microarray probes. Each GSM sample is associated with a GPL. E.g., to find out the gene symbols for a GSM, you would need to consult with the GPL used in that study.
Probe-Gene

- **Data description:**(http://www.ncbi.nlm.nih.gov/geo/)



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

Repository Browser

DataSets:	4348
Series:	105148
Platforms:	19112
Samples:	2761599

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

GSE:

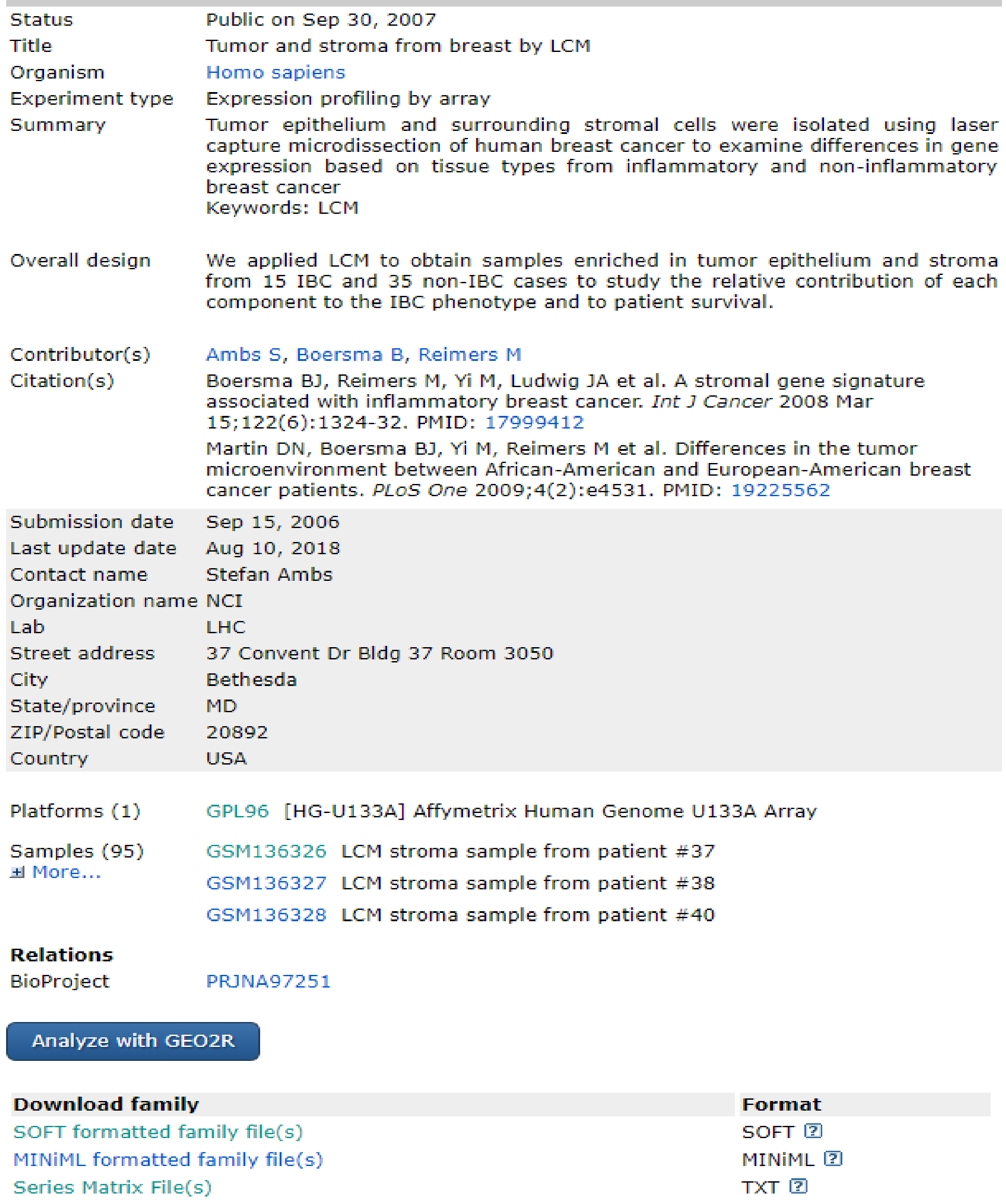
Two sections: Header and data

- **Header.Series structure**
- **Header.Samples structure**
- **!series_matrix_table**

You can download the file GSE5847_series_matrix which is already in your folder

!series_matrix_table

- **sample names (column names)**
- **probe information (row names)**
- **No gene information (GPL)**



Status	Public on Sep 30, 2007
Title	Tumor and stroma from breast by LCM
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Tumor epithelium and surrounding stromal cells were isolated using laser capture microdissection of human breast cancer to examine differences in gene expression based on tissue types from inflammatory and non-inflammatory breast cancer Keywords: LCM
Overall design	We applied LCM to obtain samples enriched in tumor epithelium and stroma from 15 IBC and 35 non-IBC cases to study the relative contribution of each component to the IBC phenotype and to patient survival.
Contributor(s)	Ambs S, Boersma B, Reimers M
Citation(s)	Boersma BJ, Reimers M, Yi M, Ludwig JA et al. A stromal gene signature associated with inflammatory breast cancer. *Int J Cancer* 2008 Mar 15;122(6):1324-32. PMID: 17999412 Martin DN, Boersma BJ, Yi M, Reimers M et al. Differences in the tumor microenvironment between African-American and European-American breast cancer patients. *PLoS One* 2009;4(2):e4531. PMID: 19225562
Submission date	Sep 15, 2006
Last update date	Aug 10, 2018
Contact name	Stefan Ambs
Organization name	NCI
Lab	LHC
Street address	37 Convent Dr Bldg 37 Room 3050
City	Bethesda
State/province	MD
ZIP/Postal code	20892
Country	USA
Platforms (1)	GPL96 [HG-U133A] Affymetrix Human Genome U133A Array
Samples (95) ⊞ More...	GSM136326 LCM stroma sample from patient #37 GSM136327 LCM stroma sample from patient #38 GSM136328 LCM stroma sample from patient #40

Relations

BioProject	PRJNA97251

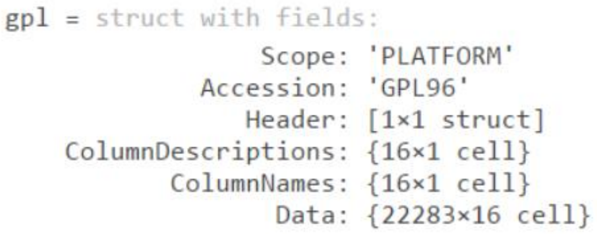
Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

GPL:

Related to the type of arrays.
Define the probes for genes

Three sections:

^PLATFORM = GPL96

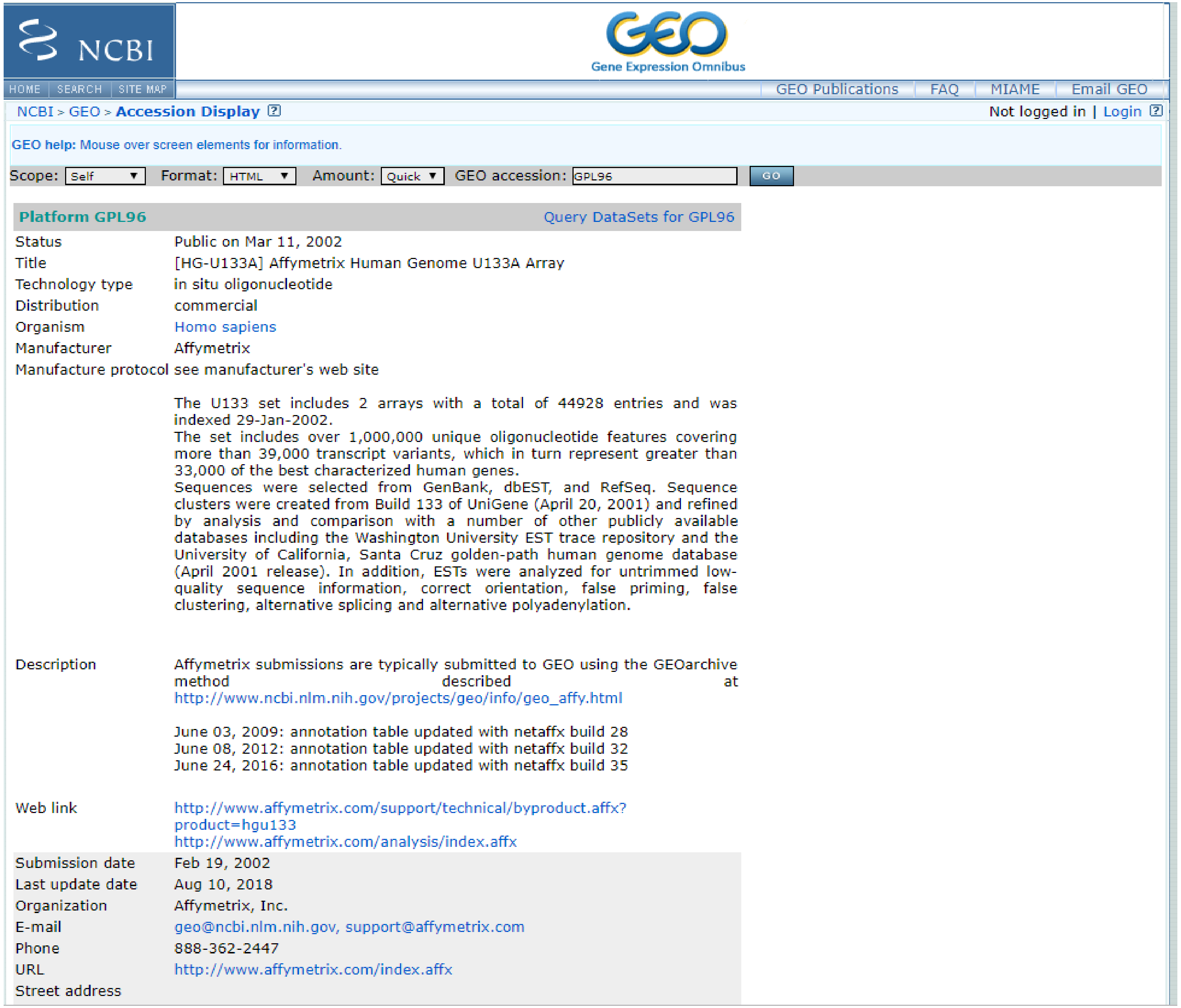


```
gpl = struct with fields:

                  Scope: 'PLATFORM'
              Accession: 'GPL96'
                 Header: [1×1 struct]
     ColumnDescriptions: {16×1 cell}
            ColumnNames: {16×1 cell}
                   Data: {22283×16 cell}
```

!platform_table_begin

We are only interested in the platform_table
Column: probes used
Row: probe name



GEO help: Mouse over screen elements for information.

Scope: Self   Format: HTML   Amount: Quick   GEO accession: GPL96   GO

Platform GPL96   Query DataSets for GPL96

Status	Public on Mar 11, 2002
Title	[HG-U133A] Affymetrix Human Genome U133A Array
Technology type	in situ oligonucleotide
Distribution	commercial
Organism	Homo sapiens
Manufacturer	Affymetrix
Manufacture protocol	see manufacturer's web site The U133 set includes 2 arrays with a total of 44928 entries and was indexed 29-Jan-2002. The set includes over 1,000,000 unique oligonucleotide features covering more than 39,000 transcript variants, which in turn represent greater than 33,000 of the best characterized human genes. Sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from Build 133 of UniGene (April 20, 2001) and refined by analysis and comparison with a number of other publicly available databases including the Washington University EST trace repository and the University of California, Santa Cruz golden-path human genome database (April 2001 release). In addition, ESTs were analyzed for untrimmed low-quality sequence information, correct orientation, false priming, false clustering, alternative splicing and alternative polyadenylation.
Description	Affymetrix submissions are typically submitted to GEO using the GEOarchive method described at http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_affy.html June 03, 2009: annotation table updated with netaffx build 28 June 08, 2012: annotation table updated with netaffx build 32 June 24, 2016: annotation table updated with netaffx build 35
Web link	http://www.affymetrix.com/support/technical/byproduct.affx?product=hgu133 http://www.affymetrix.com/analysis/index.affx
Submission date	Feb 19, 2002
Last update date	Aug 10, 2018
Organization	Affymetrix, Inc.
E-mail	geo@ncbi.nlm.nih.gov, support@affymetrix.com
Phone	888-362-2447
URL	http://www.affymetrix.com/index.affx
Street address	

bmes_downloadandparsegse5

```
url = sprintf('ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/%s/%s_series_matrix.txt.gz','GSE5847','GSE5847');
gzfile = [tempdir '/' sprintf('%s.txt.gz','GSE5847')];
fprintf('Downloading %s ...\n',url);
urlwrite(url, gzfile);
files = gunzip( gzfile );
file = files{1};
fprintf('Reading %s ...\n',file);
gsedata = geoseriesread( file );
```

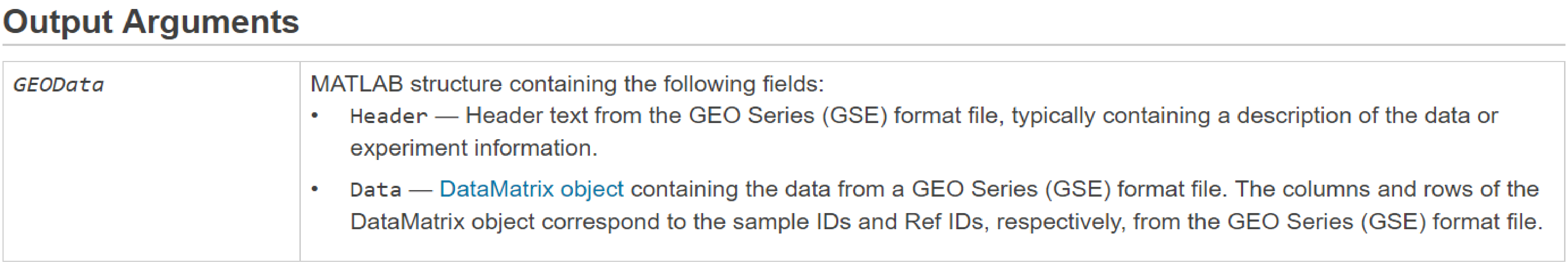
Geoseriesread bioinformatics toolbox

Read Gene Expression Omnibus (GEO) Series (GSE) format data

https://www.mathworks.com/help/bioinfo/ref/geoseriesread.html

Syntax

GEOData = geoseriesread(*File*)



Output Arguments

`GEOData`	MATLAB structure containing the following fields: • `Header` — Header text from the GEO Series (GSE) format file, typically containing a description of the data or experiment information. • `Data` — DataMatrix object containing the data from a GEO Series (GSE) format file. The columns and rows of the DataMatrix object correspond to the sample IDs and Ref IDs, respectively, from the GEO Series (GSE) format file.

bmes_downloadandparsegpl7

```
url = sprintf('https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?form=text&acc=%s&view=full',gplid);
file = [tempdir '/' 'GPL96' '.txt'];
urlwrite(url, file);
gpldata = geosoftread( file );
```

geosoftread

Read Gene Expression Omnibus (GEO) SOFT format data

https://www.mathworks.com/help/bioinfo/ref/geosoftread.html

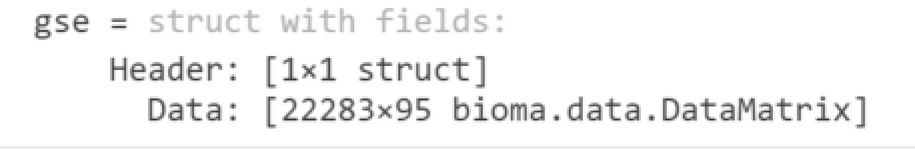
Syntax

GEOSOFTData = geosoftread(*File*)

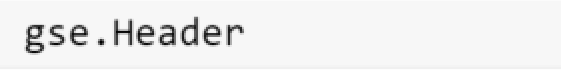
GEOSOFTData	MATLAB structure containing information from a GEO SOFT format file.

Fields	Description
Scope	Type of file read (SAMPLE, DATASET, or PLATFORM)
Accession	Accession number for record in GEO database.
Header	Microarray experiment information.
ColumnDescriptions	Cell array containing descriptions of columns in the data.
ColumnNames	**Cell array containing names of columns in the data.**
Data	**Array containing microarray data.**
Identifier (GDS files only)	Cell array containing probe IDs.
IDRef (GDS files only)	Cell array containing indices to probes.

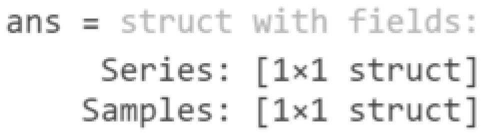
GSE: Data Structures After Downloading



```
gse = struct with fields:
    Header: [1×1 struct]
      Data: [22283×95 bioma.data.DataMatrix]
```



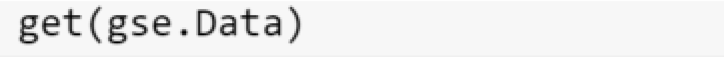
```
gse.Header
```



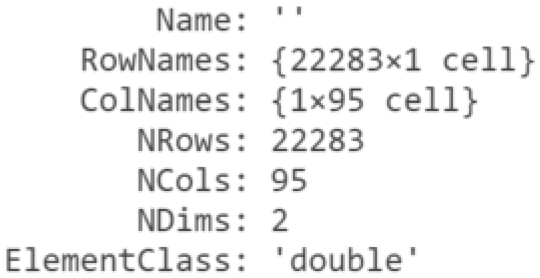
```
ans = struct with fields:
     Series: [1×1 struct]
    Samples: [1×1 struct]
```

Gse.data. DataMatrix object, similar to a Matlab table but with row and column names
You need to follow the rules of DataMatrix object

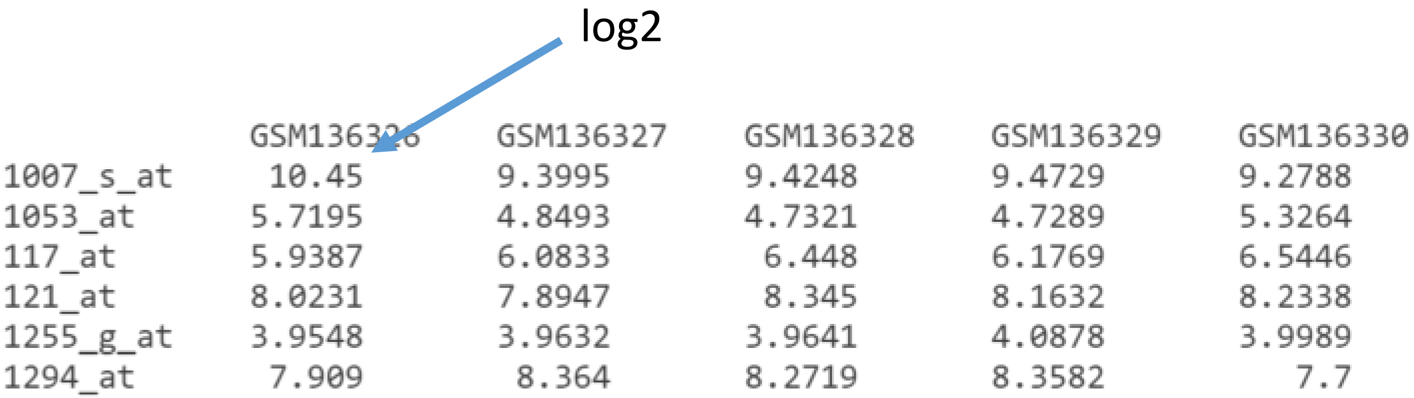
d=



```
get(gse.Data)
```

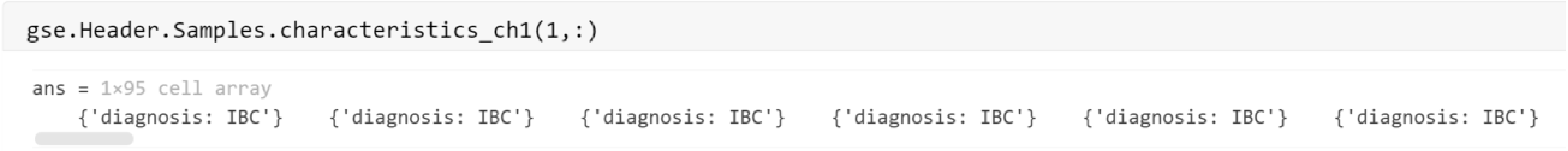


```
           Name: ''
       RowNames: {22283×1 cell}
       ColNames: {1×95 cell}
          NRows: 22283
          NCols: 95
          NDims: 2
   ElementClass: 'double'
```



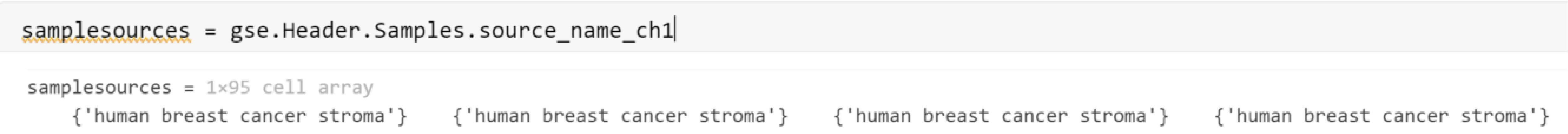
log2

```
             GSM136326    GSM136327    GSM136328    GSM136329    GSM136330
1007_s_at     10.45       9.3995       9.4248       9.4729       9.2788
1053_at      5.7195       4.8493       4.7321       4.7289       5.3264
117_at       5.9387       6.0833        6.448       6.1769       6.5446
121_at       8.0231       7.8947        8.345       8.1632       8.2338
1255_g_at    3.9548       3.9632       3.9641       4.0878       3.9989
1294_at       7.909        8.364       8.2719       8.3582          7.7
```



```
gse.Header.Samples.characteristics_ch1(1,:)
```

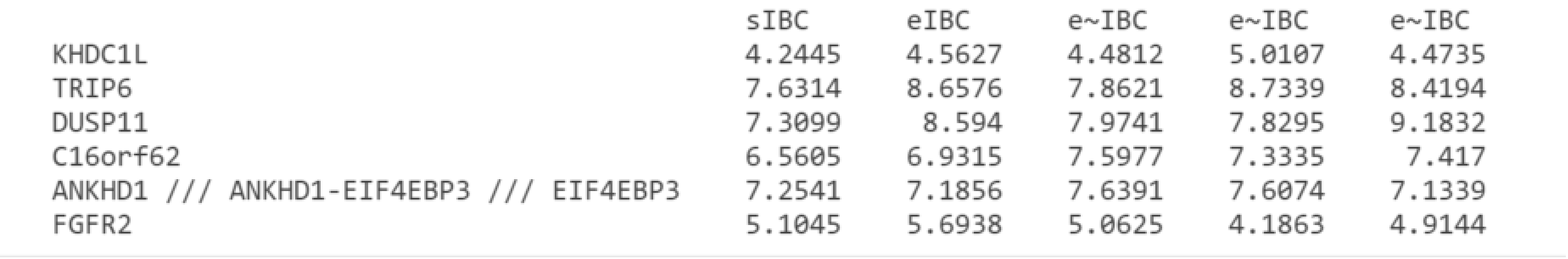
```
ans = 1×95 cell array
    {'diagnosis: IBC'}    {'diagnosis: IBC'}    {'diagnosis: IBC'}    {'diagnosis: IBC'}    {'diagnosis: IBC'}    {'diagnosis: IBC'}
```



```
samplesources = gse.Header.Samples.source_name_ch1
```

```
samplesources = 1×95 cell array
    {'human breast cancer stroma'}    {'human breast cancer stroma'}    {'human breast cancer stroma'}    {'human breast cancer stroma'}
```

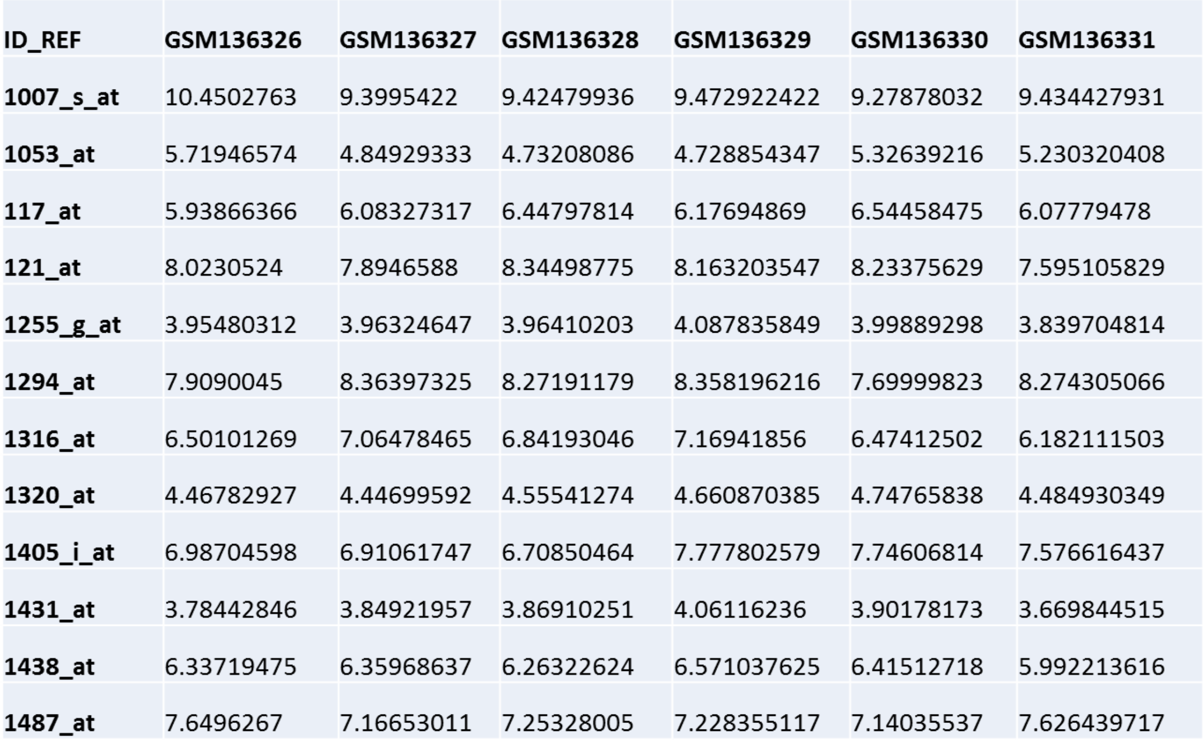
d: after changing column and row names



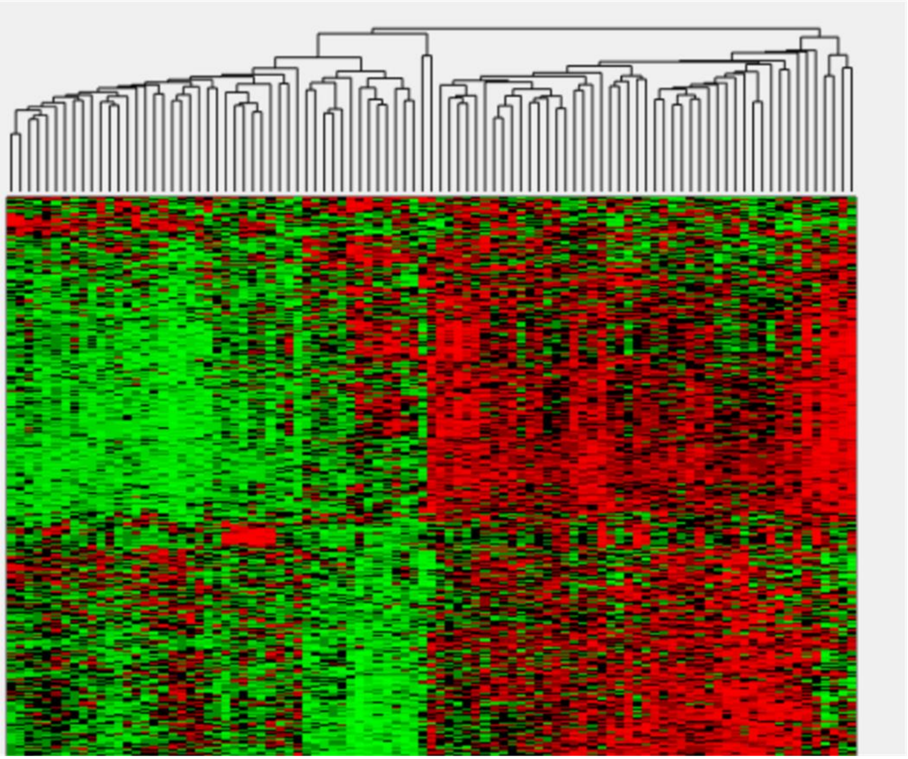
```
                                            sIBC      eIBC      e~IBC     e~IBC     e~IBC
KHDC1L                                      4.2445    4.5627    4.4812    5.0107    4.4735
TRIP6                                       7.6314    8.6576    7.8621    8.7339    8.4194
DUSP11                                      7.3099     8.594    7.9741    7.8295    9.1832
C16orf62                                    6.5605    6.9315    7.5977    7.3335     7.417
ANKHD1 /// ANKHD1-EIF4EBP3 /// EIF4EBP3     7.2541    7.1856    7.6391    7.6074    7.1339
FGFR2                                       5.1045    5.6938    5.0625    4.1863    4.9144
```

Gene expression statistical analysis

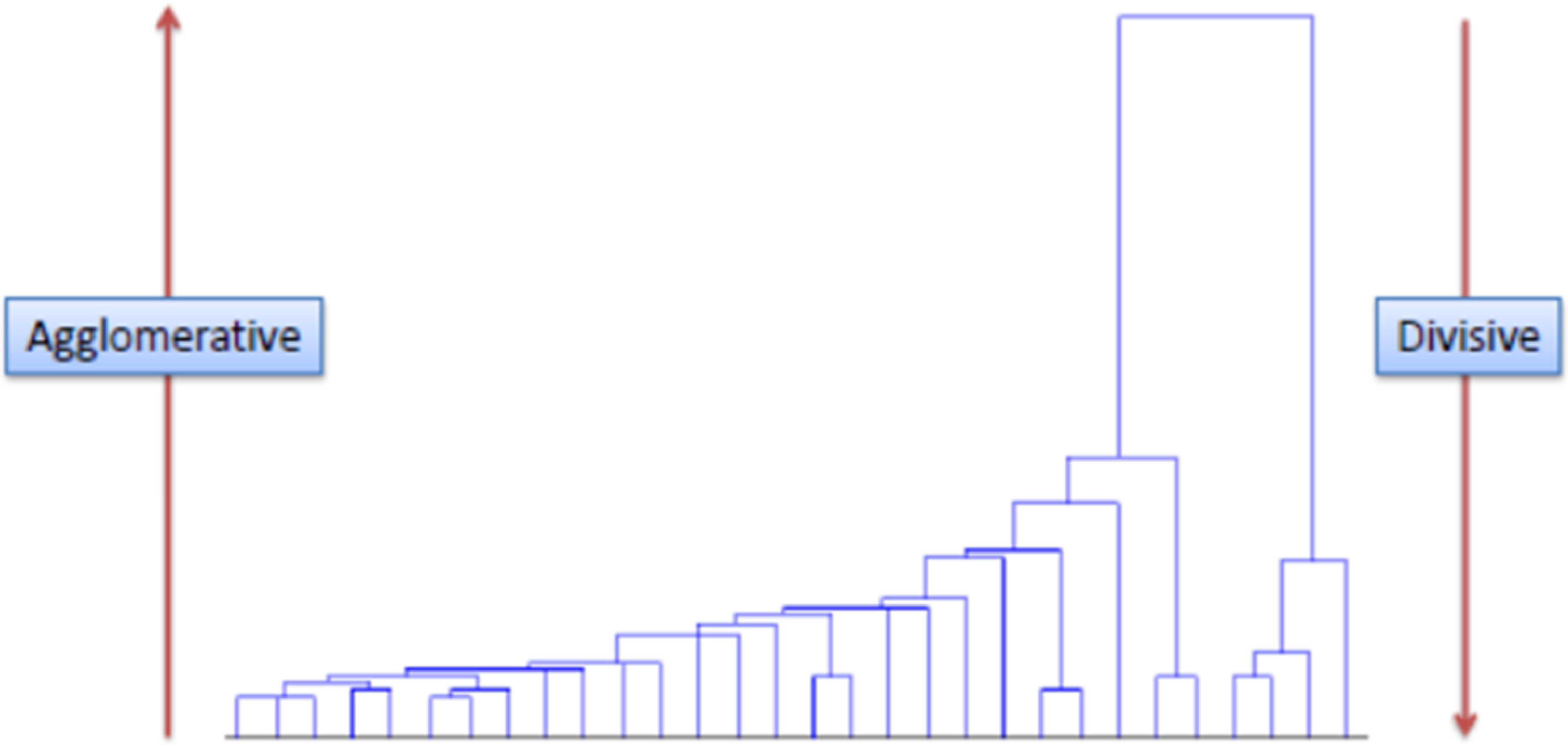
- Hierarchical Clustering
- K-means Clustering
- Principal component analysis (PCA)



ID_REF	GSM136326	GSM136327	GSM136328	GSM136329	GSM136330	GSM136331
1007_s_at	10.4502763	9.3995422	9.42479936	9.472922422	9.27878032	9.434427931
1053_at	5.71946574	4.84929333	4.73208086	4.728854347	5.32639216	5.230320408
117_at	5.93866366	6.08327317	6.44797814	6.17694869	6.54458475	6.07779478
121_at	8.0230524	7.8946588	8.34498775	8.163203547	8.23375629	7.595105829
1255_g_at	3.95480312	3.96324647	3.96410203	4.087835849	3.99889298	3.839704814
1294_at	7.9090045	8.36397325	8.27191179	8.358196216	7.69999823	8.274305066
1316_at	6.50101269	7.06478465	6.84193046	7.16941856	6.47412502	6.182111503
1320_at	4.46782927	4.44699592	4.55541274	4.660870385	4.74765838	4.484930349
1405_i_at	6.98704598	6.91061747	6.70850464	7.777802579	7.74606814	7.576616437
1431_at	3.78442846	3.84921957	3.86910251	4.06116236	3.90178173	3.669844515
1438_at	6.33719475	6.35968637	6.26322624	6.571037625	6.41512718	5.992213616
1487_at	7.6496267	7.16653011	7.25328005	7.228355117	7.14035537	7.626439717



Hierarchical Clustering



An Example of Agglomerative Clustering

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 1. Start by assigning each item to its own cluster

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Step 2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster. **BOS merged with NY**

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	0	223	1308	802	2815	2934	2786	1771
DC	223	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Step 3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster. **BOS/NY merged with DC**

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1308	802	2815	2934	2786	1771
MIA	1308	0	1329	3273	3053	2687	2037
CHI	802	1329	0	2013	2142	2054	996
SEA	2815	3273	2013	0	808	1131	1307
SF	2934	3053	2142	808	0	379	1235
LA	2786	2687	2054	1131	379	0	1059
DEN	1771	2037	996	1307	1235	1059	0

LA and SF become the closet, they will merge and for a new cluster

Step 3. Compute distances (similarities) between the new cluster and each of the old clusters. **Then LA merged with SF**

	BOS/NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

The cluster of BOS/NYC/DC is the closest to CHI

Repeat Step 2 and 3

LA/SF merged with SEA

	BOS/NY/DC/CHI	MIA	SEA	SF/LA	DEN
BOS/NY/DC/CHI	0	1075	2013	2054	996
MIA	1075	0	3273	2687	2037
SEA	2013	3273	0	808	1307
SF/LA	2054	2687	808	0	1059
DEN	996	2037	1307	1059	0

BOS/NY/DC/CHI/ merged with DEN

	BOS/NY/DC/CHI	MIA	SF/LA/SEA	DEN
BOS/NY/DC/CHI	0	1075	2013	996
MIA	1075	0	2687	2037
SF/LA/SEA	2054	2687	0	1059
DEN	996	2037	1059	0

BOS/NY/DC/CHI/DEN merged with SF/LA/SEA

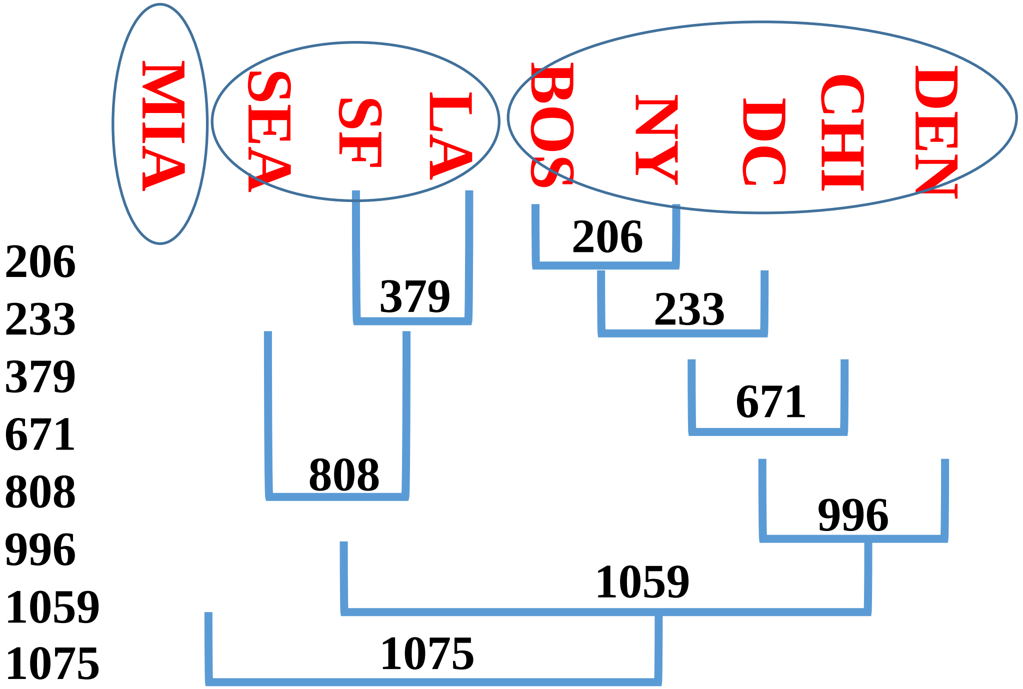
	BOS/NY/DC/CHI/DEN	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

BOS/NY/DC/CHI/DEN/SF/LA/SEA merged with MIA

	BOS/NY/DC/CHI/DEN/SF/LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

Now trace back to the clustering process plot the dendrogram

1. BOS merged with NY
2. BOS/NY merged with DC
3. LA merged with SF and LA/SF merged with SEA
4. BOS/NY/DC/CHI/ merged with DEN
5. BOS/NY/DC/CHI/DEN merged with SF/LA/SEA
6. BOS/NY/DC/CHI/DEN/SF/LA/SEA merged with MIA



What is k-Means Cluster Analysis?

k-means cluster analysis is an algorithm that groups similar objects into groups called clusters.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

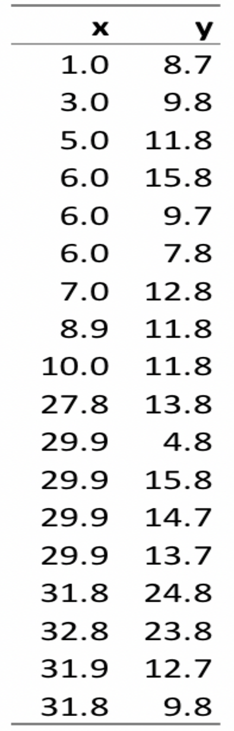
The K-means algorithm identifies *k* number of centroids (a centroid is the imaginary or real location representing the center of the cluster), and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

How *k*-means cluster analysis works

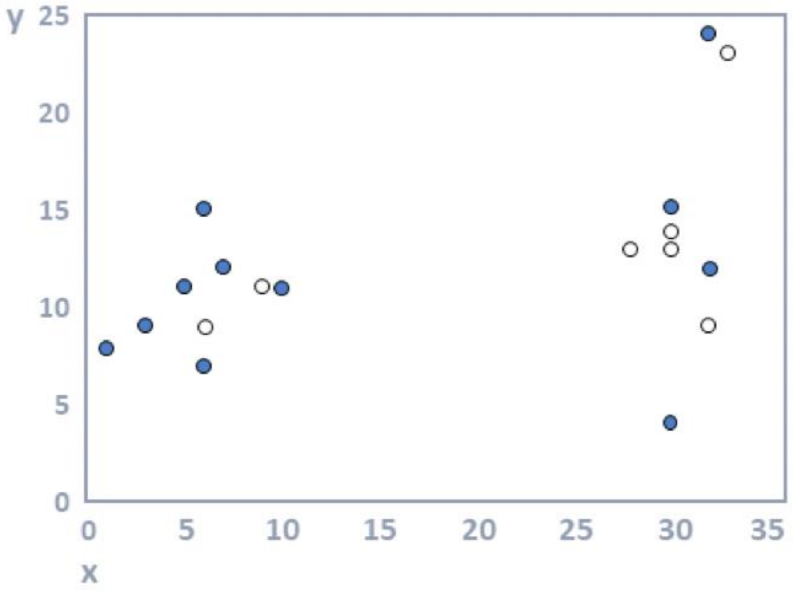
Step 1: Specify the number of clusters (*k*). The first step in *k*-means is to specify the number of clusters, which is referred to as *k*. Traditionally researchers will conduct *k*-means multiple times, exploring different numbers of clusters (e.g., from 2 through 10).

Step 2: Allocate objects to clusters. The most straightforward approach is to randomly assign objects to clusters, but there are many other approaches (e.g., using *hierarchical clustering*).

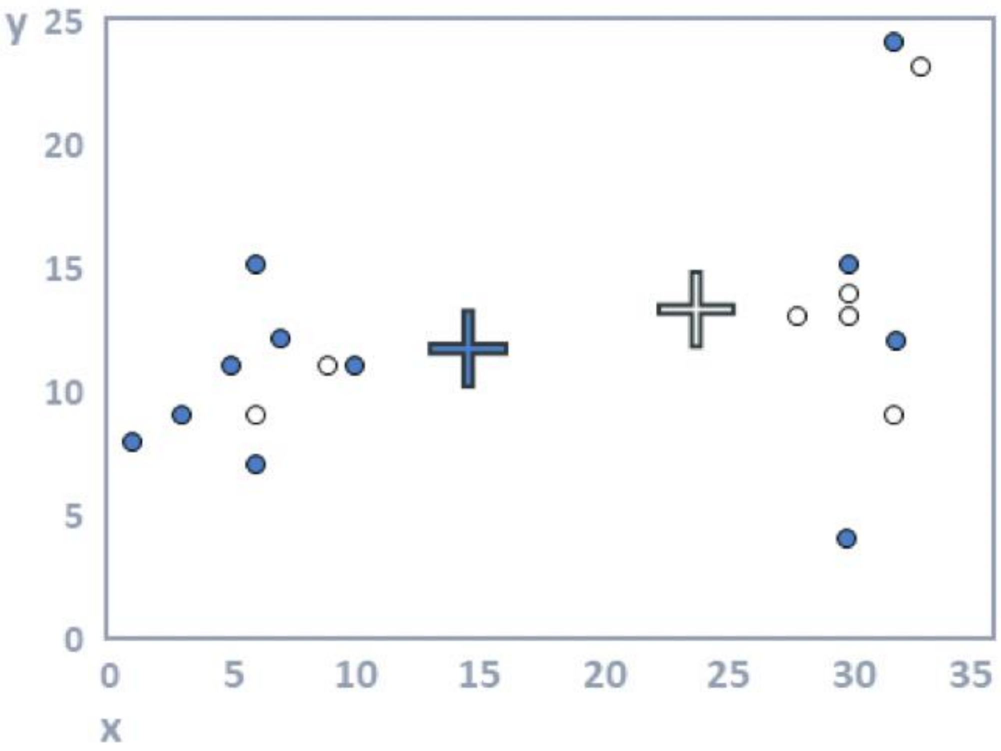
In the diagram, the 18 objects have been represented by dots on a *scatterplot,* where **x** is shown by the horizontal position of each object and **y** by the vertical. The objects have been randomly assigned to the two clusters (*k* = 2), where one cluster is shown with filled dots and the other with unfilled dots.



x	y
1.0	8.7
3.0	9.8
5.0	11.8
6.0	15.8
6.0	9.7
6.0	7.8
7.0	12.8
8.9	11.8
10.0	11.8
27.8	13.8
29.9	4.8
29.9	15.8
29.9	14.7
29.9	13.7
31.8	24.8
32.8	23.8
31.9	12.7
31.8	9.8

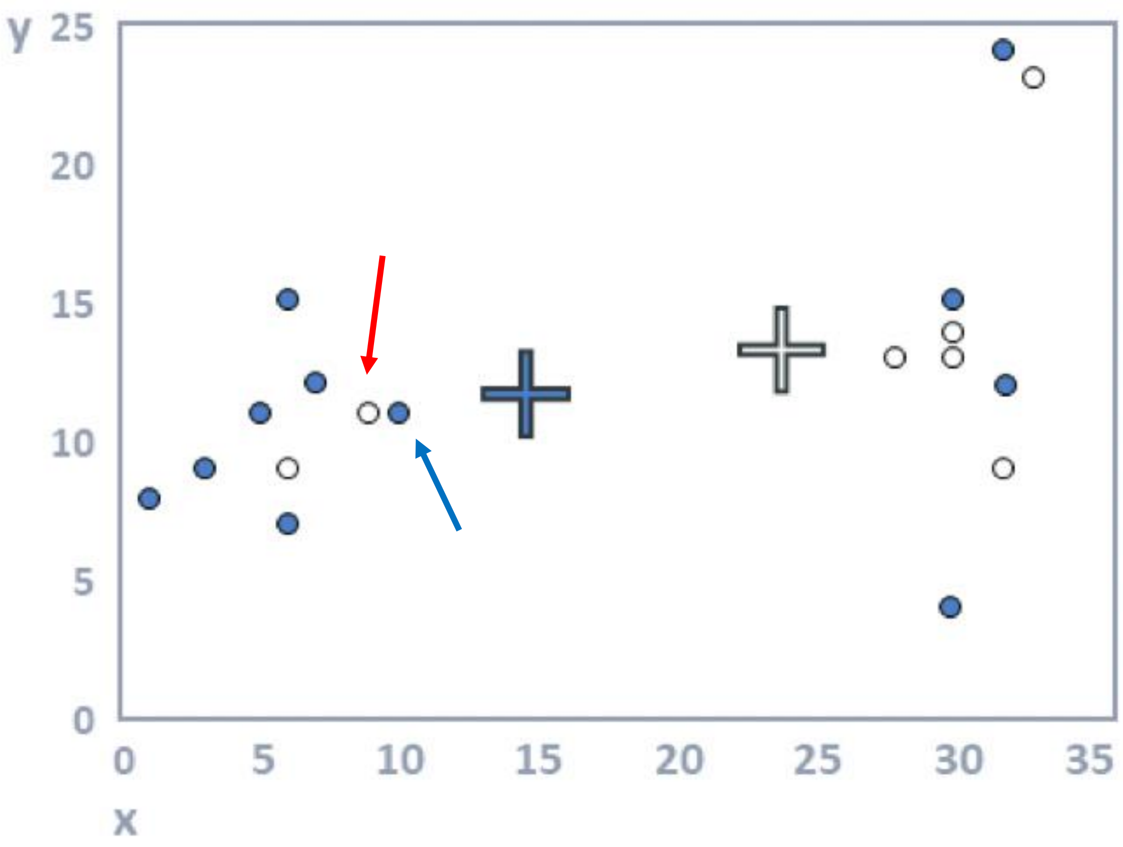


Step 3: Compute cluster means. For each cluster, the average value is computed for each of the variables. In the plot below, the average value of the filled dots for the variable represented by the horizontal position (x) of the dots is around 15; for the variable on the vertical dimension it is around twelve. These two means are represented by the filled cross. Or, stated slightly differently: the filled cross is in the middle of the black dots. Similarly, the white cross is in the middle of the white dots. These crosses are variously referred to as the *cluster centers, cluster means,* and *cluster medoids*.

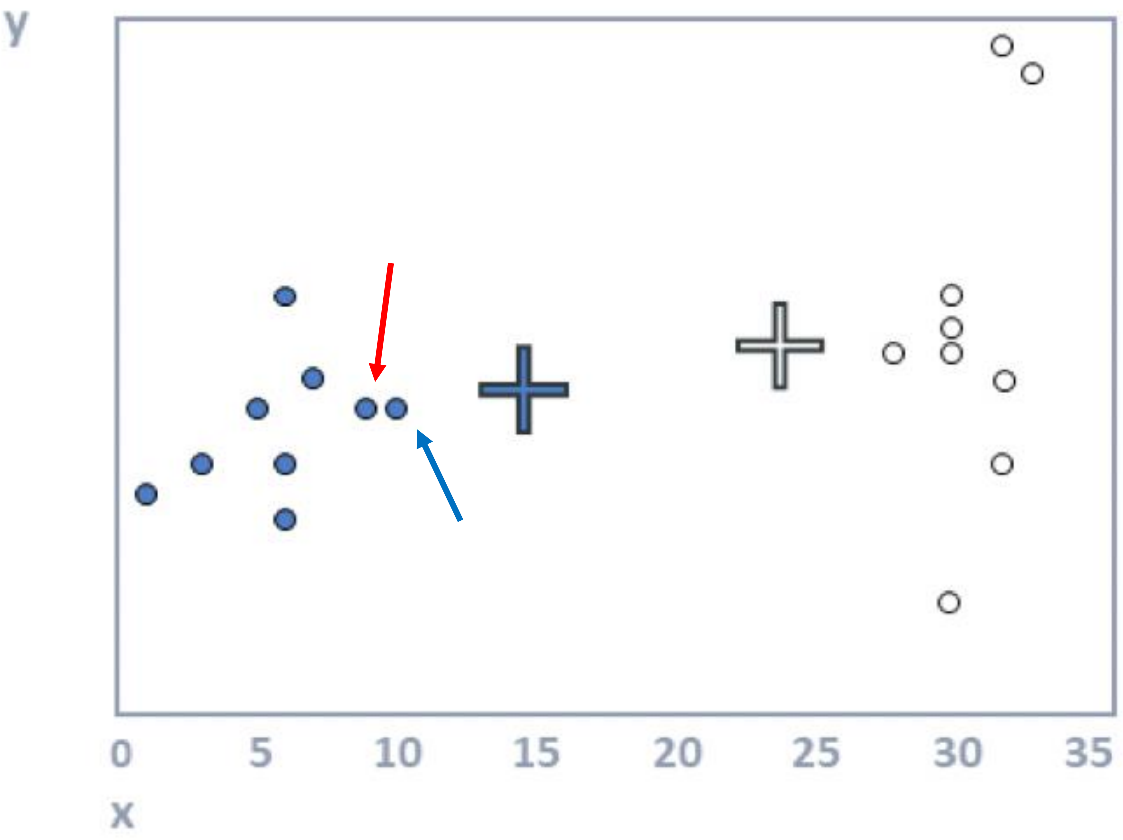


Step 4: Allocate each observation to the closest cluster center. In the plot above, some of the filled dots are closer to the white cross and some of the white dots are closer to the black cross. When we reallocate the observations to the closest clusters we get the plot below.

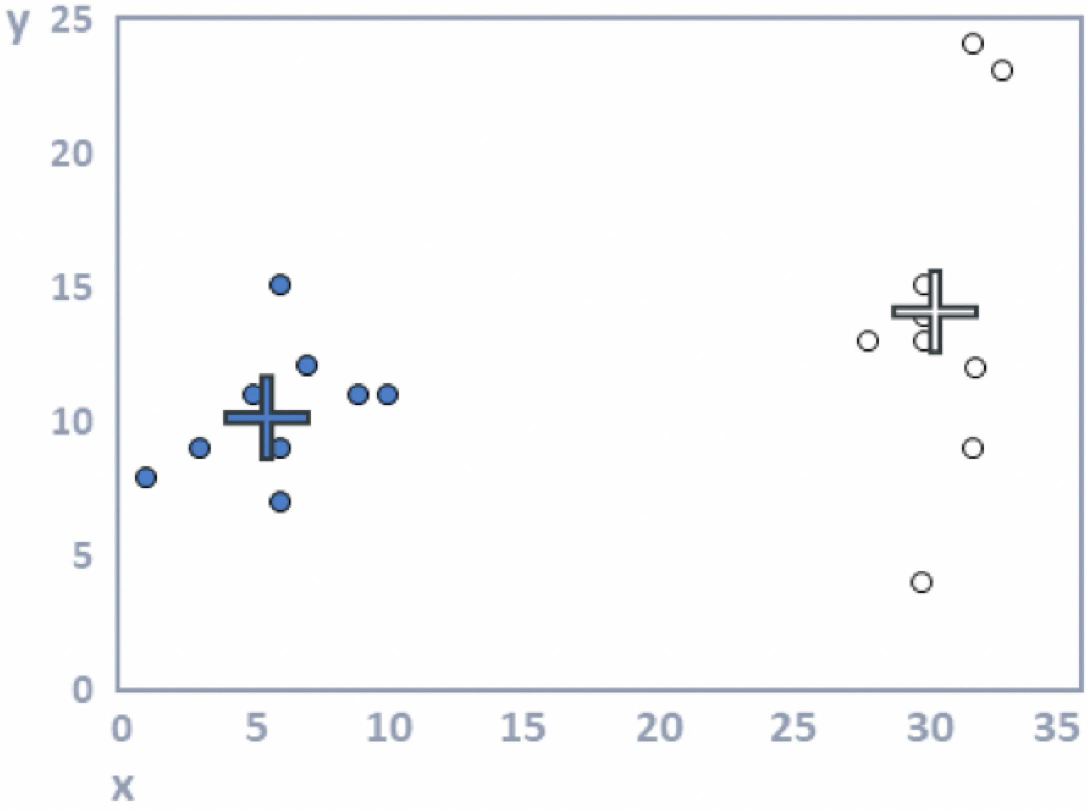
Randomly assigned clusters



New clusters

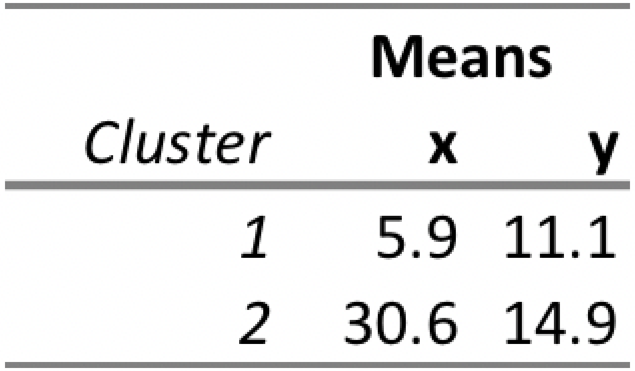


Step 5: Repeat steps 3 and 4 until the solution converges. Looking at the plot above, we can see that the crosses (the cluster means) are no longer accurate. In the following plot they have been recomputed using step 3. In this example the cluster analysis has *converged* (i.e., reallocating observations and updating means cannot improve the solution). In examples with more data a few more iterations are typically required (i.e., steps 3 and 4 are repeated until no respondents change clusters).

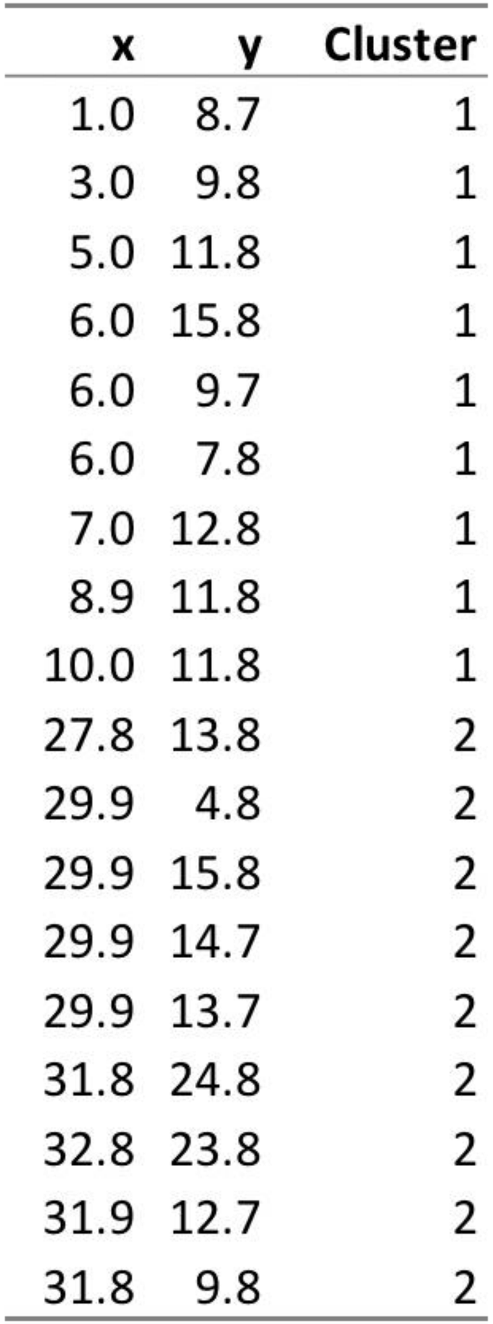


The outputs from *k*-means cluster analysis

The main output from *k*-means cluster analysis is a table showing the mean values of each cluster on the clustering variables. The *table of means* for the data examined in this article is shown below.



	Means	
Cluster	**x**	**y**
1	5.9	11.1
2	30.6	14.9



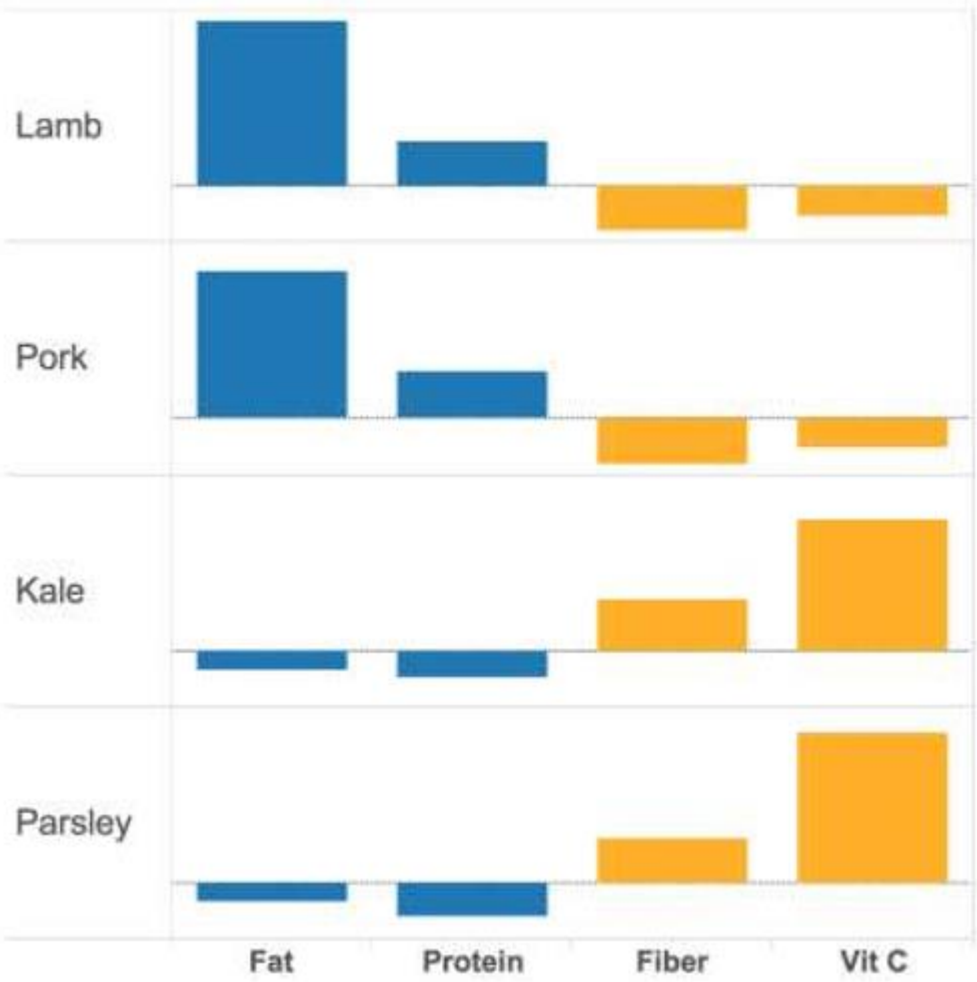
x	**y**	**Cluster**
1.0	8.7	1
3.0	9.8	1
5.0	11.8	1
6.0	15.8	1
6.0	9.7	1
6.0	7.8	1
7.0	12.8	1
8.9	11.8	1
10.0	11.8	1
27.8	13.8	2
29.9	4.8	2
29.9	15.8	2
29.9	14.7	2
29.9	13.7	2
31.8	24.8	2
32.8	23.8	2
31.9	12.7	2
31.8	9.8	2

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a simple yet powerful technique used for dimensionality reduction. Through it, we can directly decrease the number of feature variables, thereby narrowing down the important features and saving on computations.

Principal component analysis is a technique for *feature extraction* — so it combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables! *As an added benefit, each of the “new” variables after PCA are all independent of one another.*

Among four food items described by four variables (nutrients)

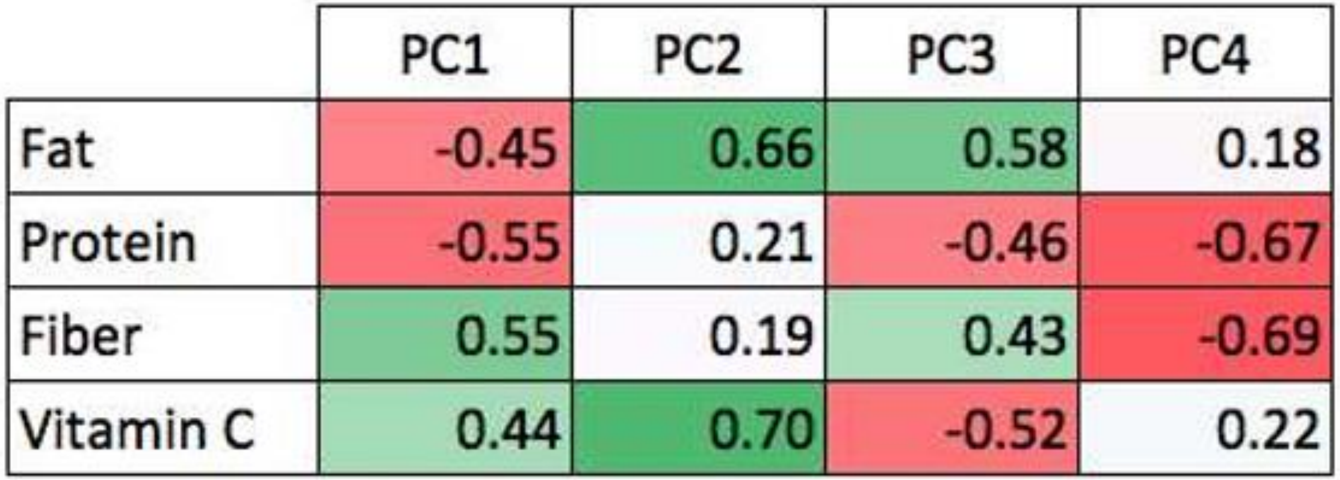


Four kinds of food with four variables (4 dimensions) based on their nutrition: Fat, Protein, Fiber and VC

We can use Fat/Protein variables to differentiate meat and vegetable. But not enough differentiate the Lamb and pork

We decide to combine the original variables (fat, protein, fiber and vc) create new independent variables to separate these four food. (PCA analysis)

PCA analysis combine correlated(or not) variables and reduce variables

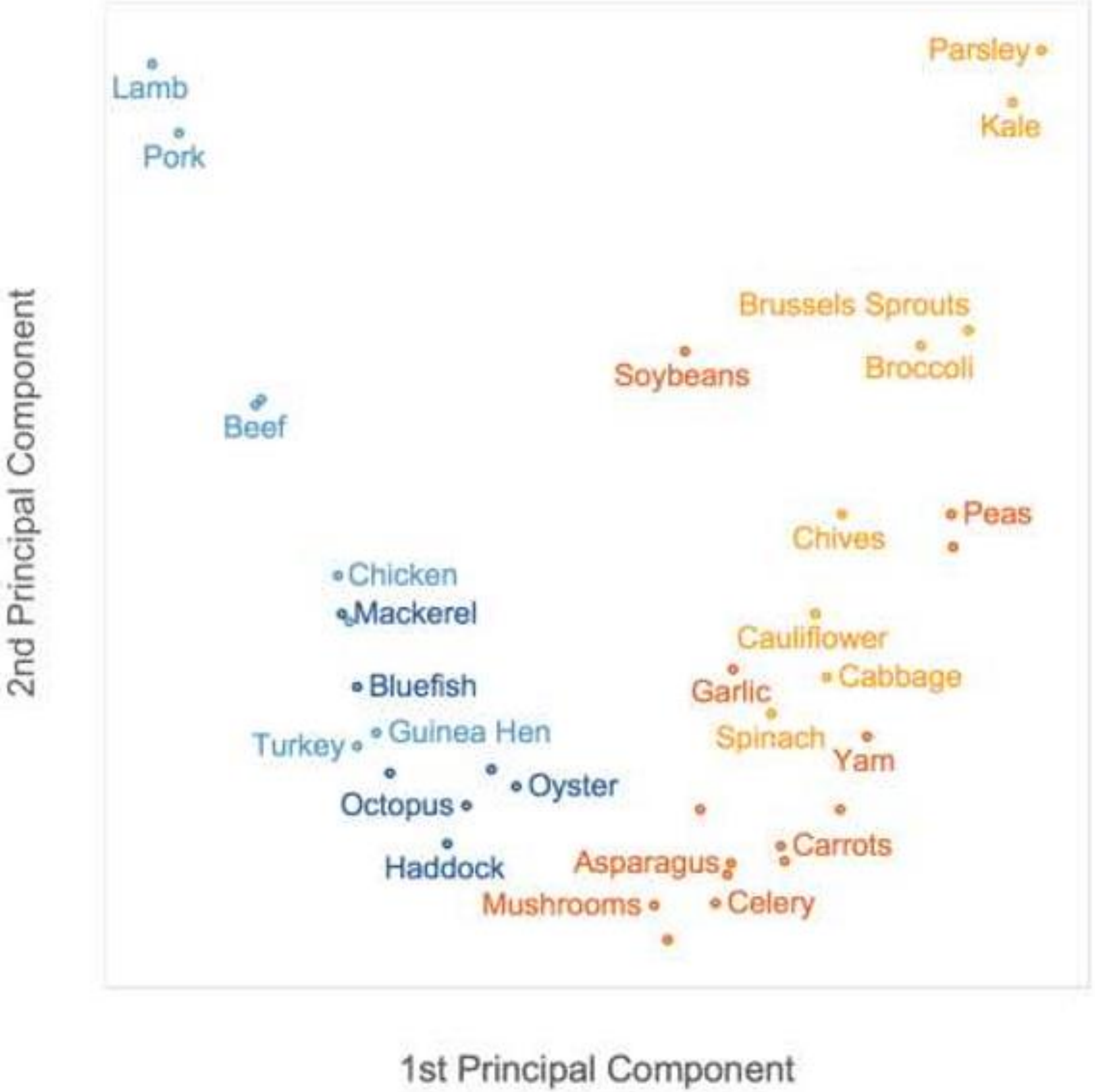


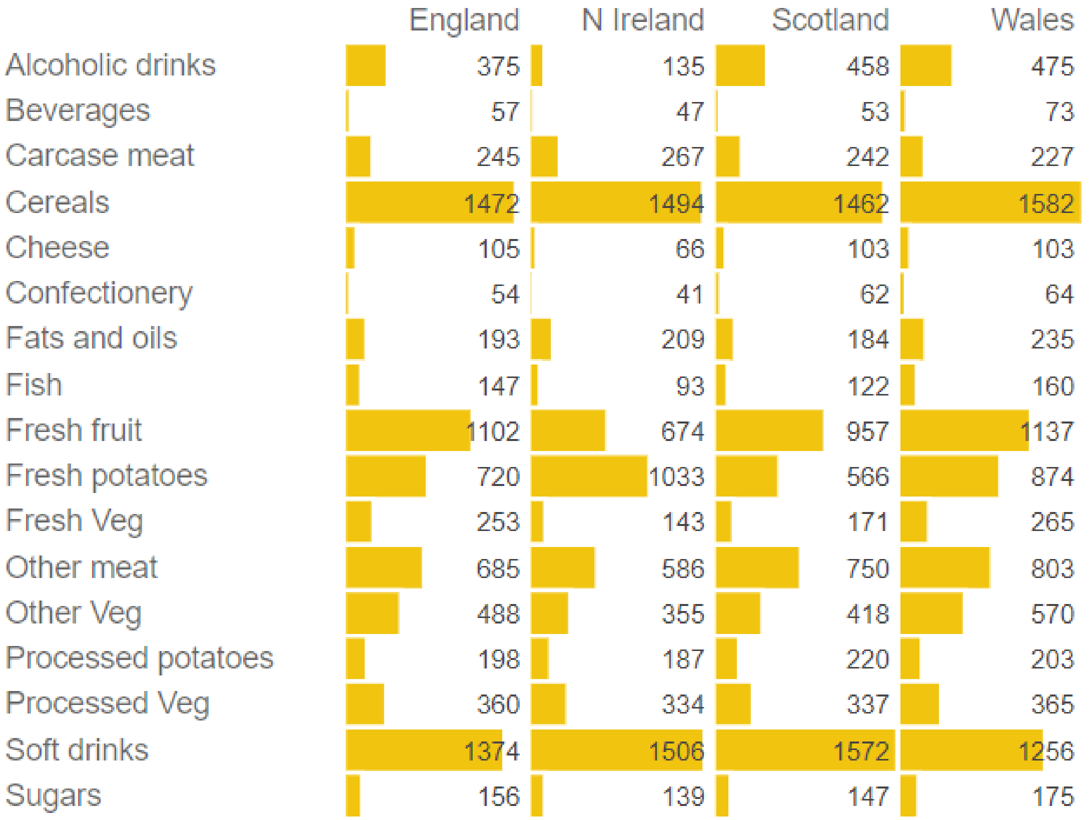
	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

In PC1, fat and protein are correlated, while fiber and vc are correlated
PC1(pork) =-0.45*Fat-0.55*Protein+0.55*Fiber+0.44*Vc

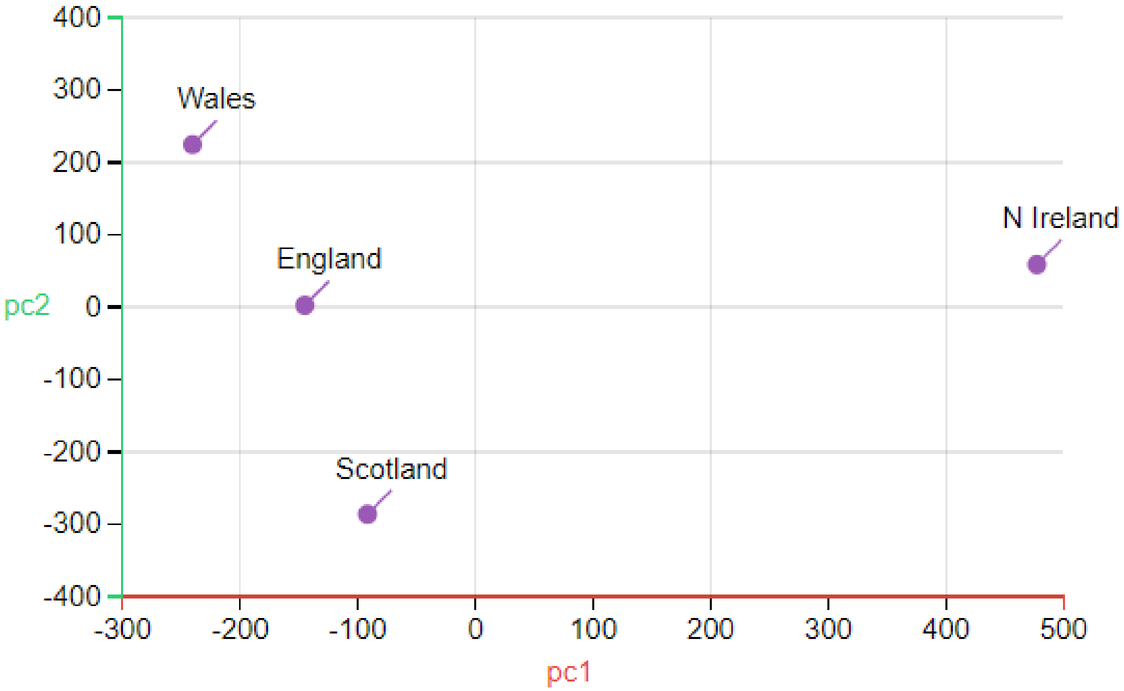
In PC2, fat and vc are correlated, while fiber and protein are correlated
PC2 (pork)=0.66*Fat+0.21*Protein+0.19*Fiber+0.70*Vc

We can calculate PC1(Lamb), PC1(parsley)…………
Plot PC1 against PC2


Lamb
Pork
Beef
Parsley
Kale
Brussels Sprouts
Soybeans
Broccoli
Chives
Peas
Chicken
Mackerel
Cauliflower
Bluefish
Garlic
Cabbage
Turkey
Guinea Hen
Spinach
Yam
Oyster
Octopus
Haddock
Asparagus
Carrots
Mushrooms
Celery
2nd Principal Component
1st Principal Component



	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



https://setosa.io/ev/principal-component-analysis/

1.Do you want to reduce the number of variables, but aren’t able to identify variables to completely remove from consideration?

2.Do you want to ensure your variables are independent of one another?

3.Are you comfortable making your independent variables less interpretable?