Pattern recognition for the analysis of asthma

Master Thesis
Kyriaki Nektaria Pantelidou

rijksuniversiteit groningen

University of Groningen
24/1/2019

Outline

- Introduction
- Datasets
- Preprocessing
- Biclustering
- t-SNE
- FABIA
- Results
- Conclusion
- Future work
- References

Introduction

DATA

- Huge amount of biomedical data
- Asthma datasets

METHODS

- Data analysis & Machine learning
- Unsupervised learning (biclustering)
- FABIA model

GOAL

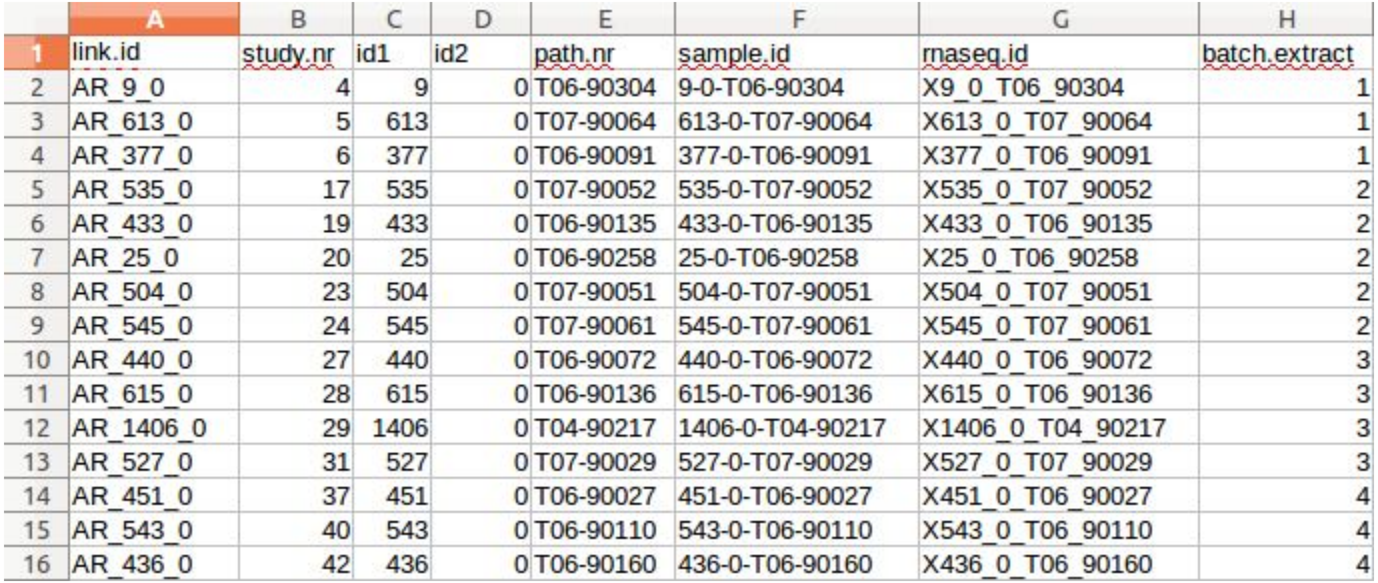
- Biomarker detection
- Discovery of patterns

Datasets

4 datasets - provided by UMCG

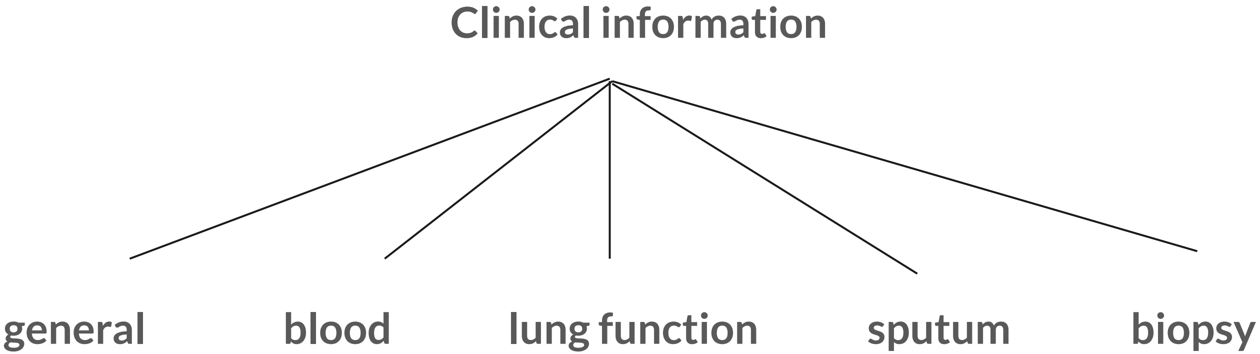
1. Clinical information for 232 subjects
2. Gene expression data for 184 subjects
3. DNA methylation data for 179 subjects
4. microRNA expression data for 206 subjects

Datasets: Clinical information



	A	B	C	D	E	F	G	H
1	link.id	study.nr	id1	id2	path.nr	sample.id	rnaseq.id	batch.extract
2	AR_9_0	4	9	0	T06-90304	9-0-T06-90304	X9_0_T06_90304	1
3	AR_613_0	5	613	0	T07-90064	613-0-T07-90064	X613_0_T07_90064	1
4	AR_377_0	6	377	0	T06-90091	377-0-T06-90091	X377_0_T06_90091	1
5	AR_535_0	17	535	0	T07-90052	535-0-T07-90052	X535_0_T07_90052	2
6	AR_433_0	19	433	0	T06-90135	433-0-T06-90135	X433_0_T06_90135	2
7	AR_25_0	20	25	0	T06-90258	25-0-T06-90258	X25_0_T06_90258	2
8	AR_504_0	23	504	0	T07-90051	504-0-T07-90051	X504_0_T07_90051	2
9	AR_545_0	24	545	0	T07-90061	545-0-T07-90061	X545_0_T07_90061	2
10	AR_440_0	27	440	0	T06-90072	440-0-T06-90072	X440_0_T06_90072	3
11	AR_615_0	28	615	0	T06-90136	615-0-T06-90136	X615_0_T06_90136	3
12	AR_1406_0	29	1406	0	T04-90217	1406-0-T04-90217	X1406_0_T04_90217	3
13	AR_527_0	31	527	0	T07-90029	527-0-T07-90029	X527_0_T07_90029	3
14	AR_451_0	37	451	0	T06-90027	451-0-T06-90027	X451_0_T06_90027	4
15	AR_543_0	40	543	0	T06-90110	543-0-T06-90110	X543_0_T06_90110	4
16	AR_436_0	42	436	0	T06-90160	436-0-T06-90160	X436_0_T06_90160	4

Datasets: Clinical information



Clinical information

general

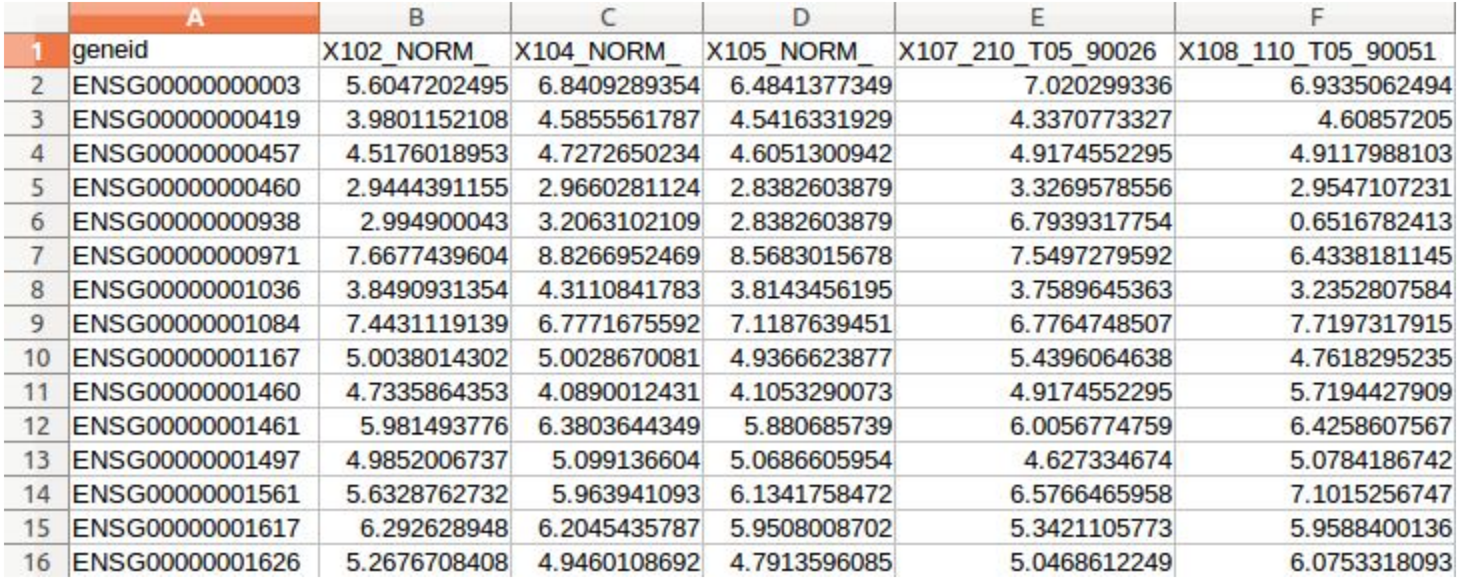
blood

lung function

sputum

biopsy

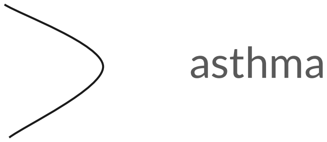
Datasets: Gene expression



	A	B	C	D	E	F
1	geneid	X102_NORM_	X104_NORM_	X105_NORM_	X107_210_T05_90026	X108_110_T05_90051
2	ENSG00000000003	5.6047202495	6.8409289354	6.4841377349	7.020299336	6.9335062494
3	ENSG00000000419	3.9801152108	4.5855561787	4.5416331929	4.3370773327	4.60857205
4	ENSG00000000457	4.5176018953	4.7272650234	4.6051300942	4.9174552295	4.9117988103
5	ENSG00000000460	2.9444391155	2.9660281124	2.8382603879	3.3269578556	2.9547107231
6	ENSG00000000938	2.994900043	3.2063102109	2.8382603879	6.7939317754	0.6516782413
7	ENSG00000000971	7.6677439604	8.8266952469	8.5683015678	7.5497279592	6.4338181145
8	ENSG00000001036	3.8490931354	4.3110841783	3.8143456195	3.7589645363	3.2352807584
9	ENSG00000001084	7.4431119139	6.7771675592	7.1187639451	6.7764748507	7.7197317915
10	ENSG00000001167	5.0038014302	5.0028670081	4.9366623877	5.4396064638	4.7618295235
11	ENSG00000001460	4.7335864353	4.0890012431	4.1053290073	4.9174552295	5.7194427909
12	ENSG00000001461	5.981493776	6.3803644349	5.880685739	6.0056774759	6.4258607567
13	ENSG00000001497	4.9852006737	5.099136604	5.0686605954	4.627334674	5.0784186742
14	ENSG00000001561	5.6328762732	5.963941093	6.1341758472	6.5766465958	7.1015256747
15	ENSG00000001617	6.292628948	6.2045435787	5.9508008702	5.3421105773	5.9588400136
16	ENSG00000001626	5.2676708408	4.9460108692	4.7913596085	5.0468612249	6.0753318093

Datasets: Labels

- **asthma**
  - current asthma
  - clinical remission
  - complete remission
  - healthy



asthma

- **currentsmoking**
  - smoker
  - non smoker
- **ics.use**
  - ics
  - no ics

Preprocessing

Clinical information dataset

- Numeric categories
- Elimination of missing values
  - Delete columns with NaN values count > 35%
  - Delete rows with at least 1 NaN value

- Normalization
  - z-score transformation

Biclustering

“Biclustering groups both rows and columns of a matrix simultaneously”

“ A bicluster is a pair of a row (gene) set and a column (sample) set for which the rows are similar to each other on the columns and vice versa”

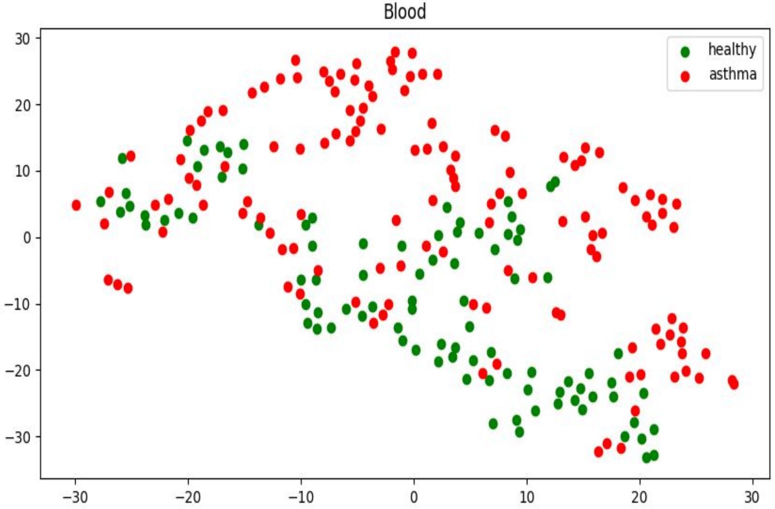
- overlapping biclusters
- part of the matrix

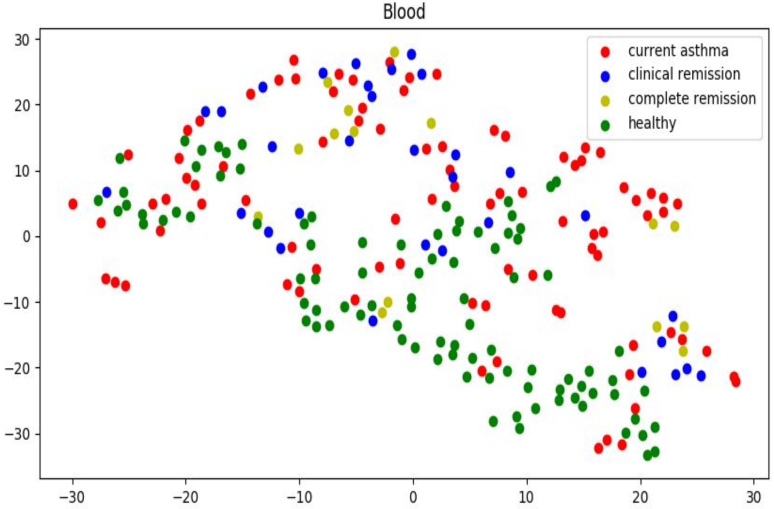
t-SNE

- Dimensionality reduction
- Visualization in 2 dimensions
- General view of the datasets
- Application to each dataset individually

t-SNE: Clinical information dataset

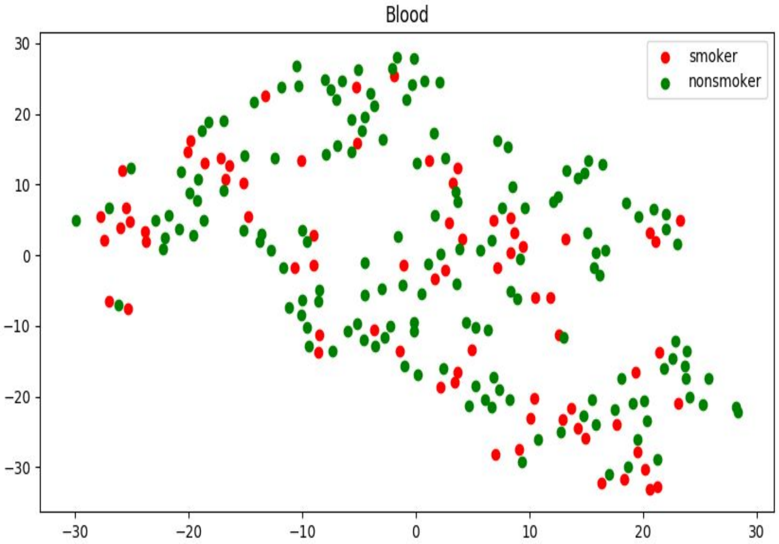
Blood

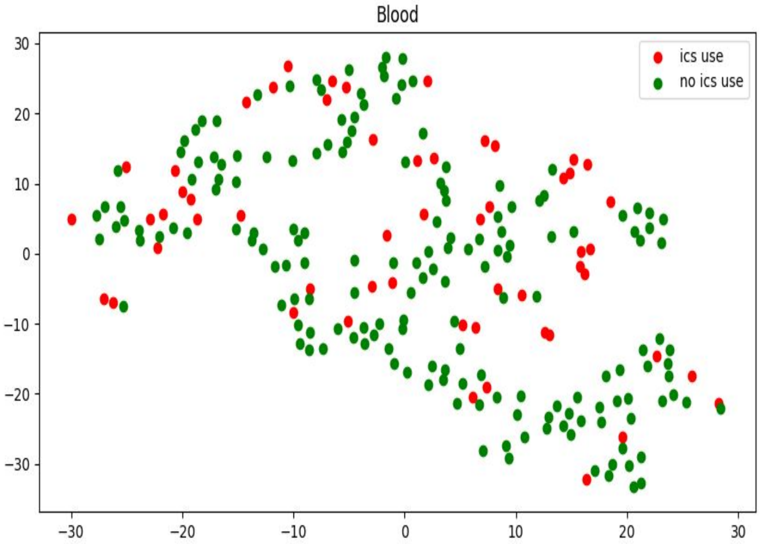




t-SNE: Clinical information dataset

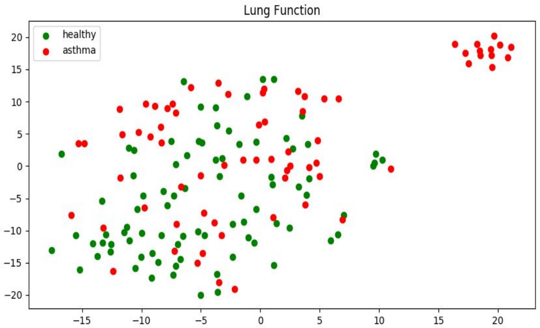
Blood



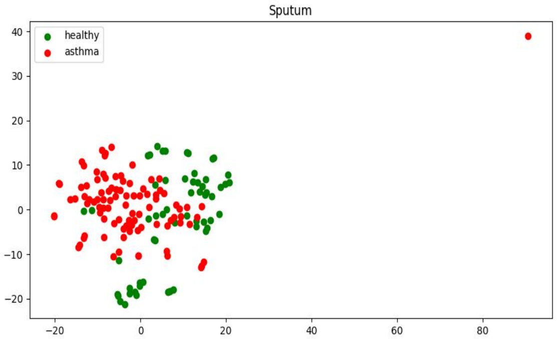


t-SNE: Clinical information dataset

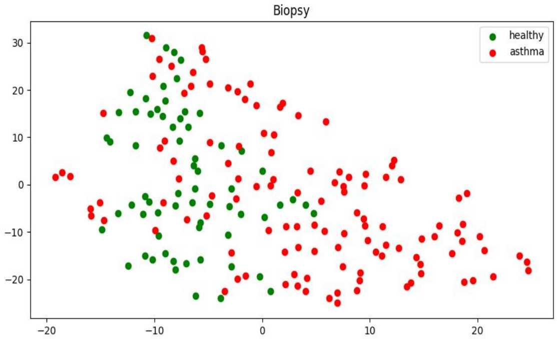
Lung Function



Sputum

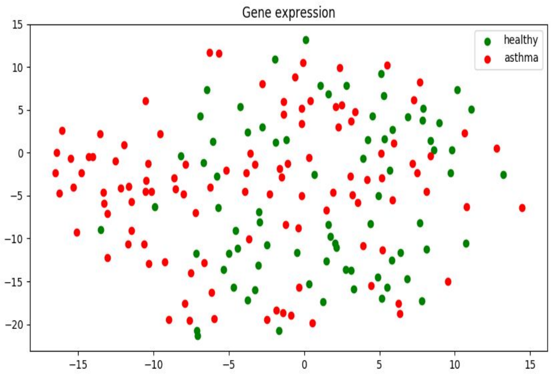


Biopsy

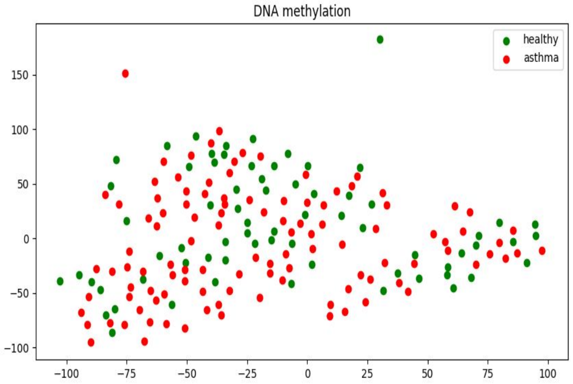


t-SNE: rest dataset

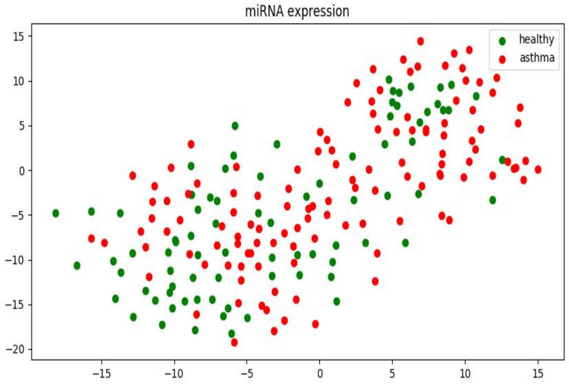
gene expression



DNA methylation



microRNA



FABIA: Factor Analysis for Bicluster Acquisition

- Multiplicative
- Generative
- Linear dependencies between genes & conditions

$$X = \sum_{i=1}^{p} \lambda_i \, z_i^T + \Upsilon = \Lambda \, Z + \Upsilon$$

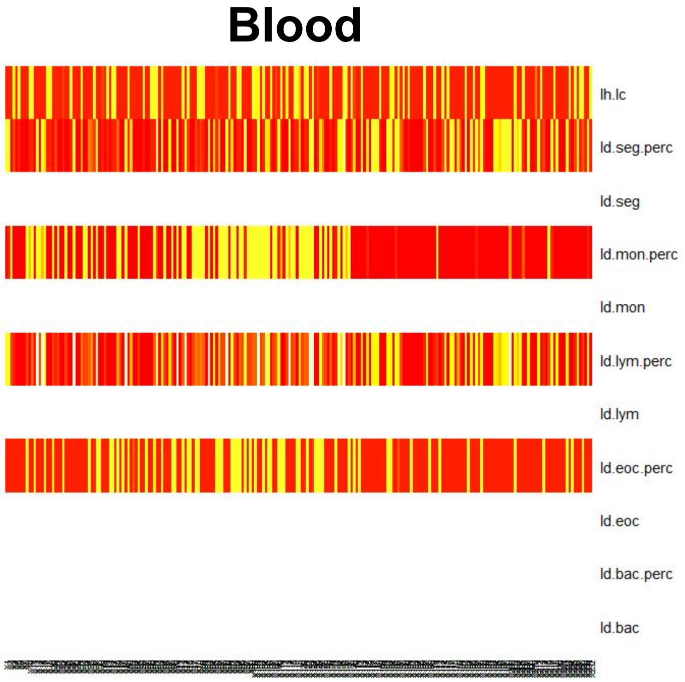
$$\overbrace{\begin{bmatrix}0\\0\\0\\0\\1\\2\\3\\4\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}}^{\lambda} * \overbrace{\begin{bmatrix}0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\end{bmatrix}}^{z^T} = \overbrace{\begin{bmatrix}0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&1&2&3&4&5&0&0&0\\0&0&0&0&0&0&0&2&4&6&8&10&0&0&0\\0&0&0&0&0&0&0&3&6&9&12&15&0&0&0\\0&0&0&0&0&0&0&4&8&12&16&20&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\end{bmatrix}}^{\lambda \, z^T}$$

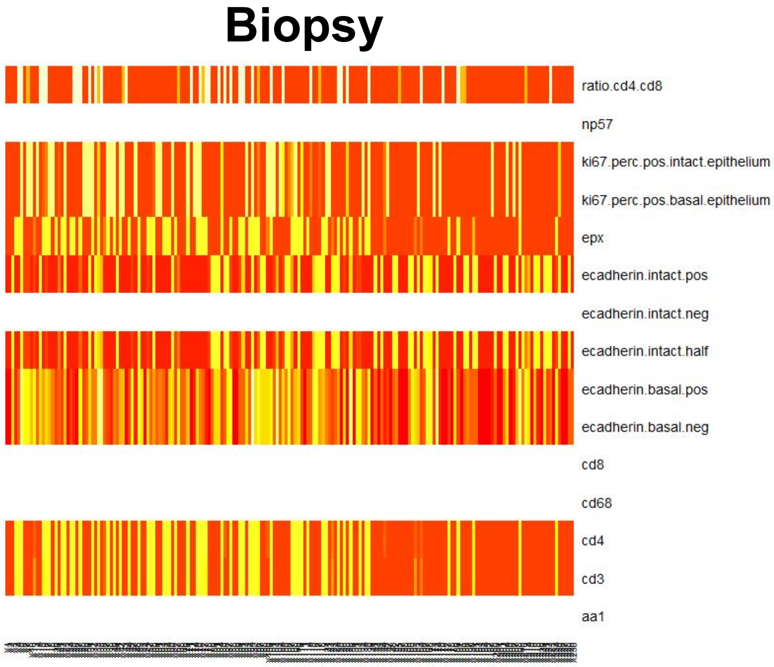
FABIA

- R implementation (package ‘fabia’)
- 4 categories
    - blood (226,11)
    - lung function (154, 51)
    - sputum (152, 33)
    - biopsy (185, 15)
- 10 runs
- counter matrices

$$\begin{array}{cc} & \text{samples} \\ \text{features} & \begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix} \end{array}$$

FABIA


Blood
lh.lc
ld.seg.perc
ld.seg
ld.mon.perc
ld.mon
ld.lym.perc
ld.lym
ld.eoc.perc
ld.eoc
ld.bac.perc
ld.bac

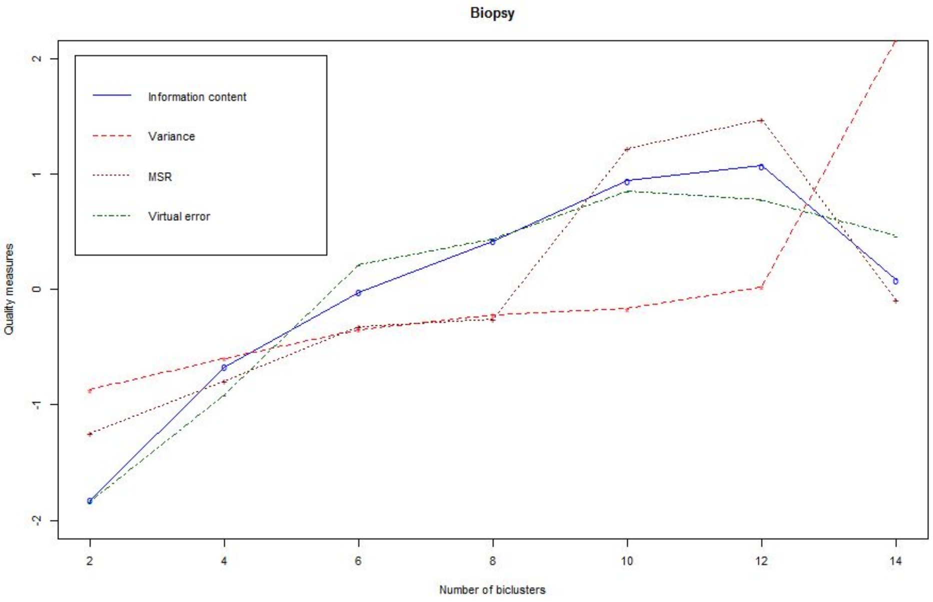

Biopsy
ratio.cd4.cd8
np57
ki67.perc.pos.intact.epithelium
ki67.perc.pos.basal.epithelium
epx
ecadherin.intact.pos
ecadherin.intact.neg
ecadherin.intact.half
ecadherin.basal.pos
ecadherin.basal.neg
cd8
cd68
cd4
cd3
aa1

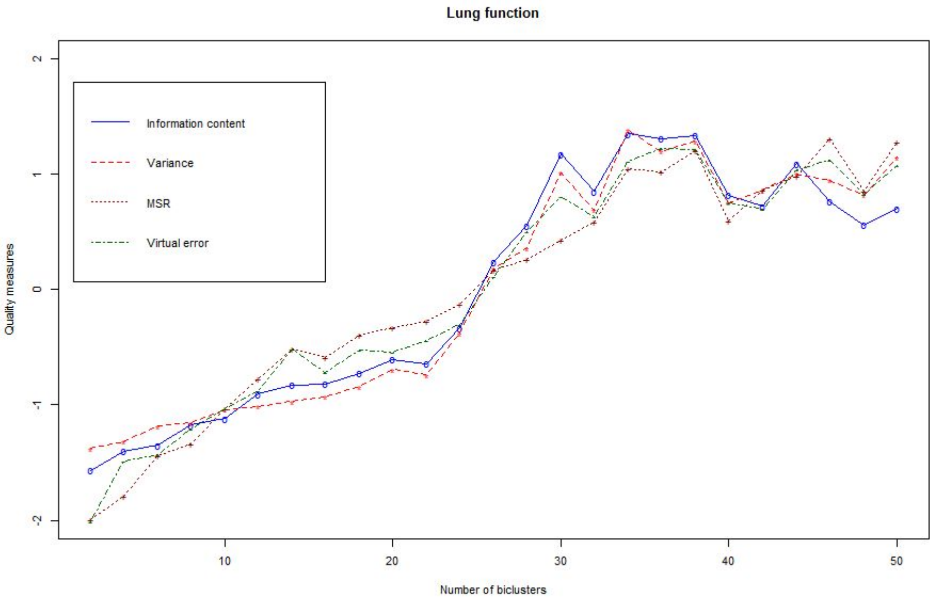
FABIA: Bicluster evaluation

Quality measures:

- Information content
- Variance
- Mean Squared Residue (MSR)
- Virtual error

FABIA: Bicluster evaluation





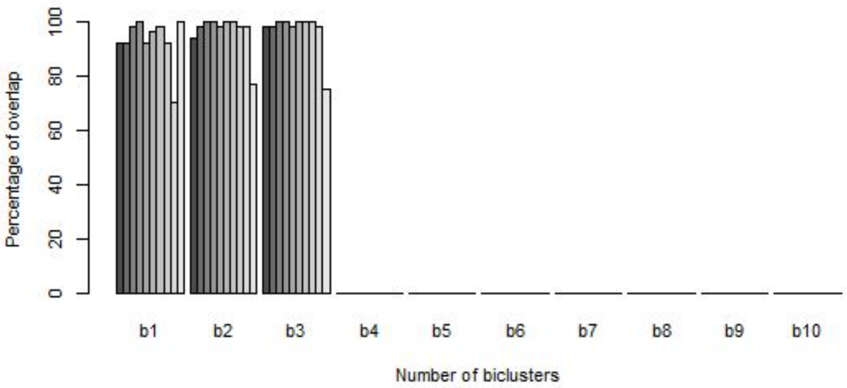
FABIA

For every category:

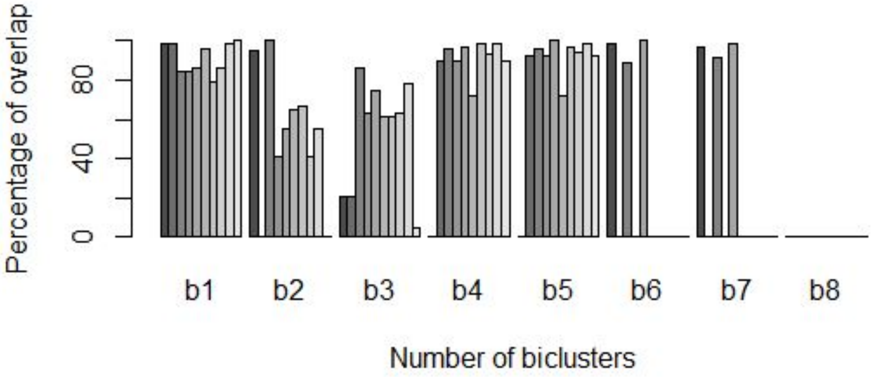
- 10 runs
- Keep the most robust biclusters (based on percentage of overlap > 80%)

FABIA

Blood

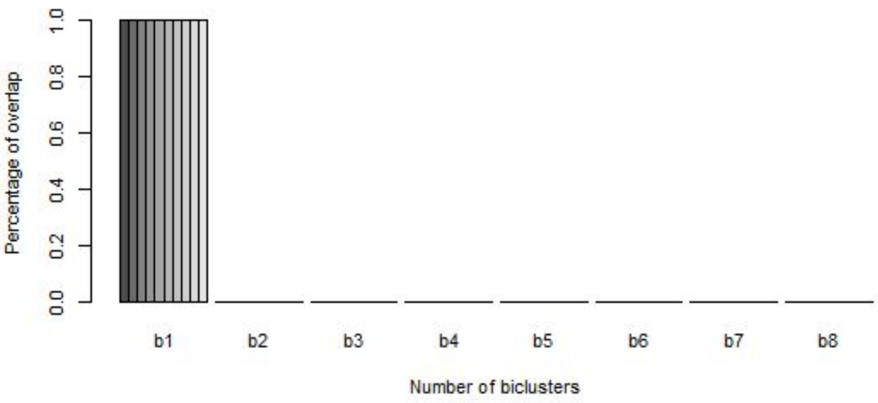


Lung Function

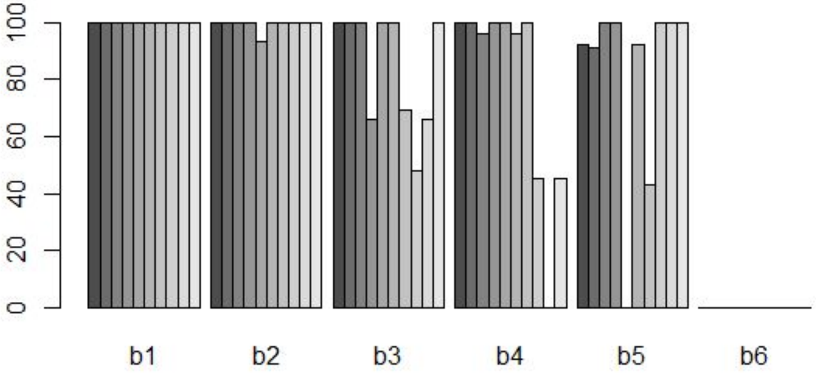


FABIA

Sputum



Biopsy



Results: Blood

1. Bicluster 1
   a. ld.lym.perc
   b. ld.seg.perc

   → 35 asthma 31 healthy

2. Bicluster 2
   a. ld.mon.perc

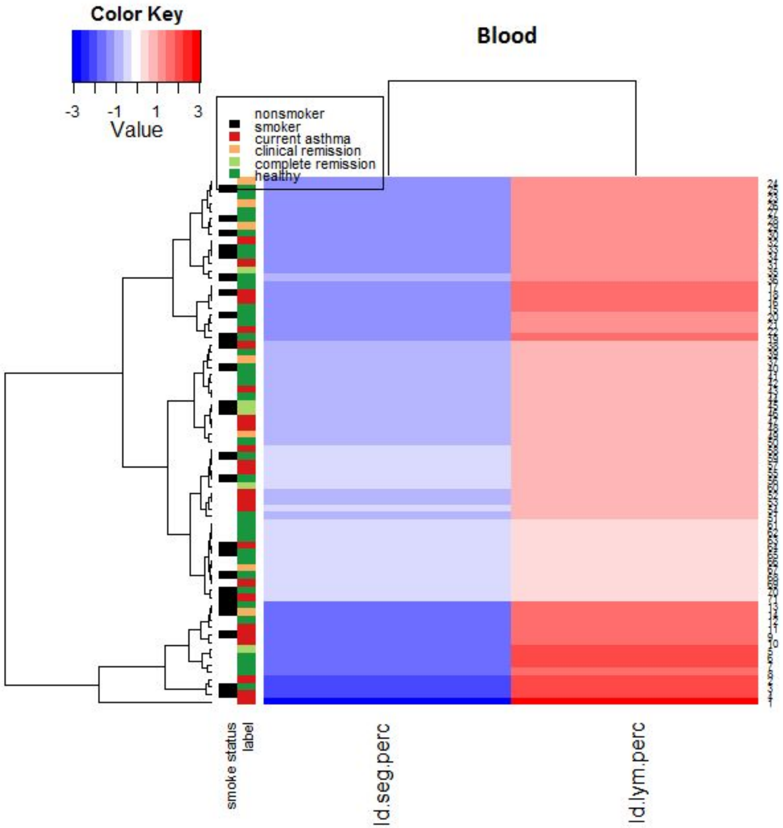
   → 69 asthma 3 healthy

3. Bicluster 3
   a. ld.eoc.perc

   → 41 asthma 14 healthy

Results: Blood

- 2 completely anticorrelated features
- No obvious pattern



Results: Lung function

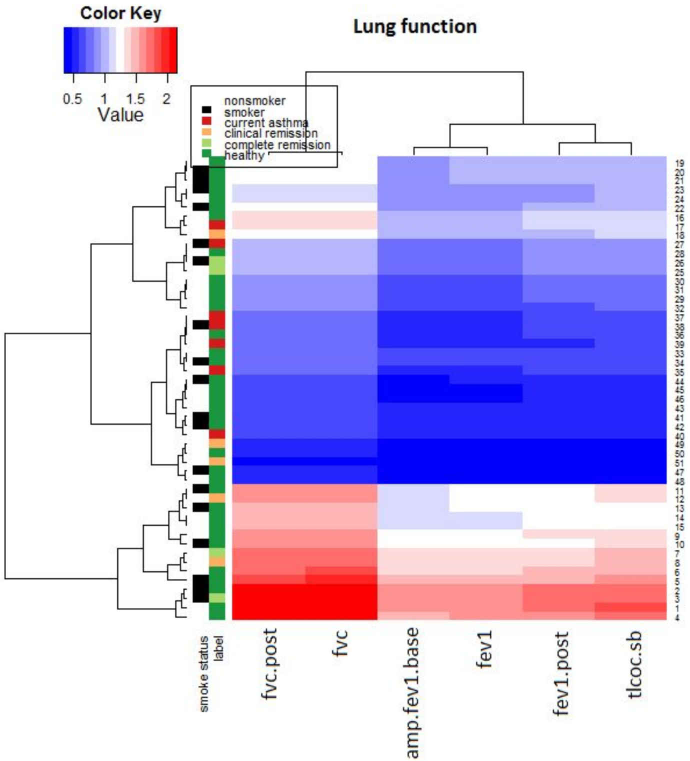
1. Bicluster 1
    a. fvc
    b. fvc.post
    c. tlcoc.sb
    d. fev1.post
    e. fev1
    f. amp.fev1.base

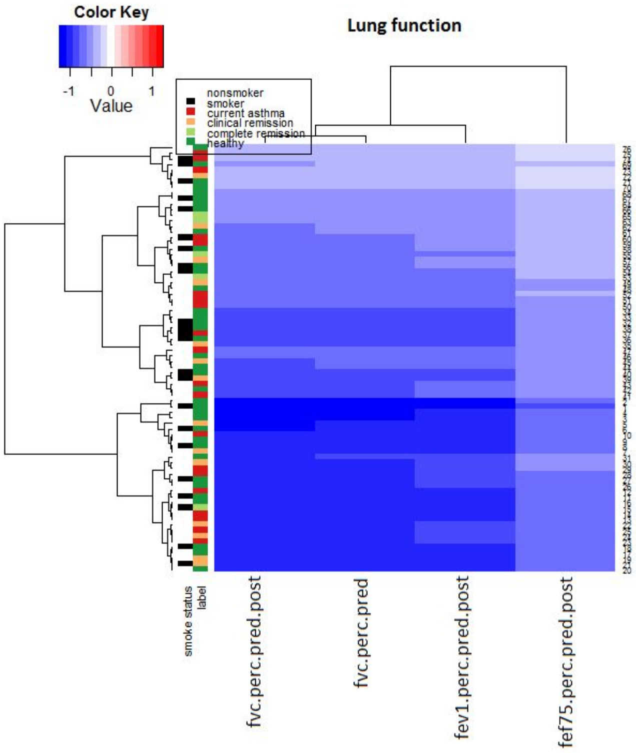
→ **19 asthma** **32 healthy**

2. Bicluster 2
    a. fvc.perc.pred.post
    b. fvc.perc.pred
    c. fev1.perc.pred.post
    d. fef75.perc.pred.post

→ **40 asthma** **37 healthy**

Results: Lung function

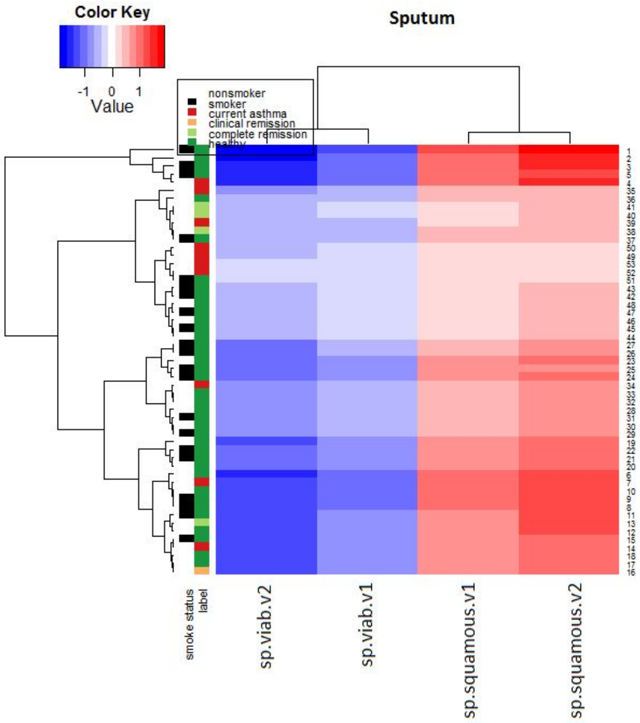




Results: Sputum

1. Bicluster 1
   a. sp.viab.v2
   b. sp.squamous.v2
   c. sp.viab.v1
   d. sp.squamous.v1

15 asthma 38 healthy



Results: Biopsy

1. Bicluster 1
   a. ecadherin.basal.pos
   b. ecadherin.basal.neg

   → **34 asthma** **12 healthy**

2. Bicluster 2
   a. ecadherin.intact.half
   b. ecadherin.intact.pos

   → **32 asthma** **27 healthy**

3. Bicluster 3
   a. ki67.perc.pos.intact.epithelium
   b. ki67.perc.pos.basal.epithelium

   → **32 asthma** **9 healthy**

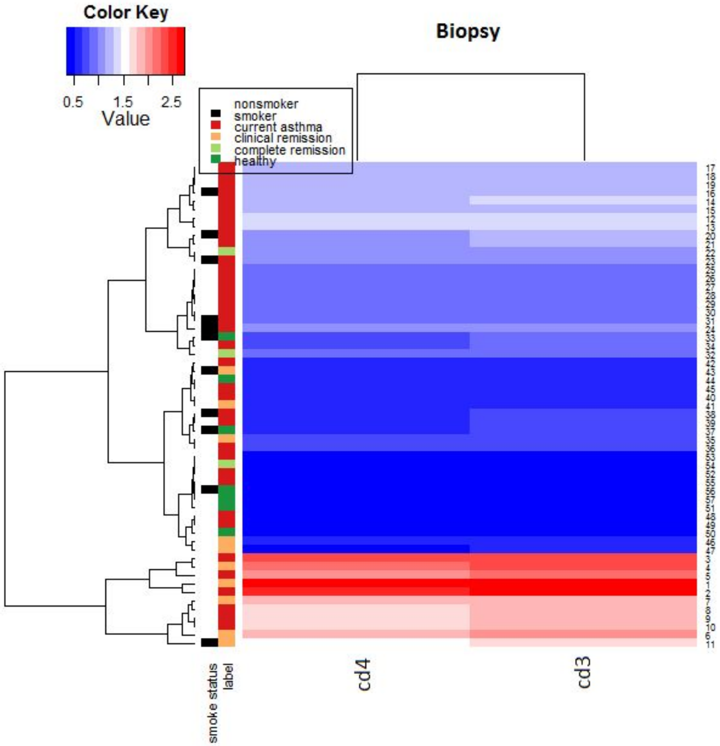
4. Bicluster 4
   a. ratio.cd4.cd8

   → **35 asthma** **6 healthy**

Results: Biopsy

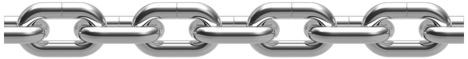
5. Bicluster 5
   a. cd3
   b. cd4

51 asthma **6 healthy**



Combination of categories

- No new combined bicluster
- Same biclusters of each category repeated



Conclusion

- Unsupervised learning is not very effective
- FABIA has limitations
- Interesting biclusters: blood & biopsy

Future work

- Work with:
    - gene expression
    - DNA methylation
    - microRNA expression
- Improved biclustering methods (e.g. GFA)
- Supervised learning

References (1)

[1] FABIA: factor analysis for bicluster acquisition, Sepp Hochreiter,, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, Luc Bijnens, Hinrich W. H. Göhlmann, Ziv Shkedy and Djork-Arné Clevert, 2010

[2] FABIA: Factor Analysis for Bicluster Acquisition — Manual for the R package, Sepp Hochreiter, 2018

[3] An effective measure for assessing the quality of biclusters, Federico Divina, Beatriz Pontes, Raul Giraldez, Jesus S.Aguilar-Ruiz, 2011

[4] Quality Measures for Gene Expression Biclusters, Beatriz Pontes, Ral Girldez, Jess S. Aguilar-Ruiz, 2015

[5] Sparse group factor analysis for biclustering of multiple data sources, Kerstin Bunte, Eemeli Leppaaho, Inka Saarinen and Samuel Kaski, 2016

References (2)

[6] Airway eosinophilia in remission and progression of asthma: Accumulation with a fast decline of FEV1, M. Broekema, F. Volbeda b,c, W. Timens, A. Dijkstra, N.A. Lee, J.J. Lee, M.E. Lodewijk, D.S. Postma, M.N. Hylkema, N.H.T. ten Hacken, 2010

[7] Persisting Remodeling and Less Airway Wall Eosinophil Activation in Complete Remission of Asthma, Martine Broekema, Wim Timens, Judith M. Vonk, Franke Volbeda, Monique E. Lodewijk, Machteld N. Hylkema, Nick H. T. ten Hacken, and Dirkje S. Postma, 2011

[8] Asthma, Stephen T. Holgate, Sally Wenzel, Dirkje S. Postma, Scott T. Weiss, Harald Renz and Peter D. Sly, 2015

Thank you!

Questions...